HIGH-DIMENSIONAL STRUCTURED REGRESSION USING CONVEX OPTIMIZATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Guo Yu

August 2018

HIGH-DIMENSIONAL STRUCTURED REGRESSION USING CONVEX OPTIMIZATION

Guo Yu, Ph.D.

Cornell University 2018

While the term "Big Data" can have multiple meanings, we consider the type of data in which the number of features can be much greater than the number of observations (also known as high-dimensional data). High-dimensional data is abundant in contemporary scientific research due to the rapid advances in new data-measurement technologies and computing power. Recent advances in statistics have witnessed great development in the field of high-dimensional data analysis. This dissertation proposes three methods that study three different components of a general framework of the high-dimensional structured regression problem. A general theme of the proposed methods is that they cast a certain structured regression as a convex optimization problem. In so doing, the theoretical properties of each method can be well studied, and efficient computation are facilitated. Each method is accompanied by thorough theoretical analysis of its performance, and also by an `R` package containing its practical implementation. We show that the proposed methods perform favorably (both theoretically and practically) compared with pre-existing methods.

BIOGRAPHICAL SKETCH

Guo Yu grew up in Yinchuan, China. His parents are both musicians, and he had never anticipated any quantitative major in his childhood. He decided to major in applied math when he was an undergraduate student in Zhejiang University, and he has never regretted this decision. In his college years, he started getting interested in statistics, where the beauty of math meets the power of computation. After graduating in 2013, Guo Yu entered the Ph.D. program in statistics at Cornell University. During his five years in the beautiful city of Ithaca, he has developed his interests in studying high-dimensional statistics problems from both theoretical and computational perspectives. In particular, he works on problems of structured sparsity, variance and covariance estimation, and large scale interaction modeling. Following the completion of his Ph.D., Guo Yu will start a postdoc position at University of Washington.

Dedicated to Mom and Dad.

ACKNOWLEDGEMENTS

I can't thank enough my advisor Jacob Bien. He is always patient with me, generous with his time, and he is such a wonderful and fun friend to be with. The past five years working with him was nothing less than my best of luck and greatest honor. I have learned so much from him: about being a statistician; about doing and presenting research; about questioning, analyzing, and diagnosing complicated problems; and about always being a curious and positive person. The list of things that he imparted me goes on and on, but what's even greater is that there is always so much more to learn from him.

Each of my committee members has substantially influenced me. Giles Hooker has influenced how I approach applied statistics through his sharp comments he shares in consulting and questions he raises in seminars. I am also deeply grateful for his support as Director of Graduate Studies. Adrian Lewis led me to the world of optimization through his wonderful lectures, and shaped how I think about linear and convex programming. I also deeply appreciate the support, funding, and advice from the faculty of the Department of Statistical Science. I especially want to thank Martin Wells, Marten Wegkamp, Jim Booth, and Beatrix Johnson.

I would also want to thank my friends and fellow classmates, especially Yuan Cheng, Zi Ye, Ze Jin, Xiaohan Yan, Rui Xu, and Yang Liu, for their support and for providing a great balance between work and fun.

Finally, I am deeply grateful to my family, Mom, and Dad, for their endless support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the development of new data-measurement technologies and computing power, high-dimensional data are ubiquitous in contemporary research fields, including biology, genetics, and information technologies. The notion of "high dimensionality" usually refers to the situation in which the number of predictors (i.e., the unknown feature parameters associated with an object) is much larger (usually in order of magnitude) than the sample size (i.e., the number of objects of interest). High-dimensionality poses a serious challenge to researchers who want to exploit informative patterns from large and complex data. Indeed, many traditional statistical methods no longer work in the presence of high dimensionality. It is very important yet challenging to build new, accurate and stable models for properly analyzing this type of data.

Recent years have witnessed great successes and advances in statistical methodology for learning high-dimensional data, both from theoretical and computational perspectives (see, e.g., Hastie et al. 2011, Bühlmann & Van De Geer 2011). A typical setting considers

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon, \tag{1.1}$$

where the whole data set consists of n independent observation pairs of (Y, X), where Y is the response variable and $X = (X_1, X_2, \cdots, X_p)$ is the predictor vector. The unobserved error (or noise) ε has mean zero and variance σ^2, and is independent of X. The unknown functional f characterizes how the p predictors relate to the response variable Y, while σ^2 captures the noise level or extent to which Y cannot be predicted from X. The high-dimensional setting corresponds to the case in which $n \ll p$. There are three different components of (1.1): (1)

the random vector X; (2) the random error ε; and (3) the function f. This thesis proposes three methods, one for modeling each of these components.

The thesis begins in Chapter 2 with a method to model local dependence structure among p predictors $X_1, \cdots, X_p$. For many known f, it is helpful to first understand how the X_j's are related to each other for better prediction of Y. In many other settings where Y is not observed, it is of primary interest to study the dependence structure among the X_j's. Many applications feature a natural ordering among elements of the random vector X. For example, $(X_1, \cdots, X_p)$ can be some variables of interest recorded over time or some genetic mutation information measured along a human chromosome. Ordered variables depend on their predecessors (in the ordering). Such structure can be characterized by a simple model, which corresponds to learning the Cholesky factor of the inverse of the covariance matrix (i.e., the precision matrix) of X. The proposed method estimates such local dependence structure by minimizing a convex penalized criterion, where the penalty is designed to induce structured sparsity that honors the ordered information in the variables.

The second component in (1.1) is the error variance σ^2, which measures the irreducible error in modeling the dependence relationship between Y and X. The problem of estimating σ^2 is actually both important and hard in many cases, and is underdeveloped compared with the vast literature in learning f. In Chapter 3 we propose two estimators of σ^2 in a setting where $f(X) = X^T\beta^*$ is the standard linear model and sparsity of β^* is assumed. The proposed estimates are remarkably simple, and they obtain statistical properties that do not depend on any assumptions on X or β^*.

The third component of the regression model (1.1) is f, i.e., the relationship

between the response Y and the predictors X. In numerous situations, additive models $f(X) = \sum_j X_j \beta_j^*$ (i.e., using only main effects X) are insufficient to predict Y. Many complex systems involve interactions among predictors, and it is important to include these interactions in f to accurately model reality. Variable selection in interaction models with a large value of p is computationally very challenging because the number of interactions grows quadratically in p. Structural assumptions are usually imposed to facilitate computation. In Chapter 4 we propose a computationally viable approach to interaction modeling without requiring any structural assumptions on the interactions. The proposed method scales well to large problem, enjoys theoretical guarantees on its performance, and compares favorably with alternative methods.

CHAPTER 2

LEARNING LOCAL DEPENDENCE IN ORDERED DATA

Portions of this chapter were published in Yu & Bien (2017b)

2.1 Introduction

Estimating large inverse covariance matrices is a fundamental problem in modern multivariate statistics. Consider a random vector $X = \left(X_1, \ldots, X_p\right)^T \in \mathbb{R}^p$ with mean zero and covariance matrix $E(XX^T) = \Sigma$. Unlike the covariance matrix, which captures marginal correlations among variables in X, the inverse covariance matrix $\Omega = \Sigma^{-1}$ (also known as the precision matrix) characterizes conditional correlations and, under a Gaussian model, $\Omega_{jk} = 0$ implies that X_j and X_k are conditionally independent given all other variables. When p is large, it is common to regularize the precision matrix estimator by making it sparse (see, e.g., Pourahmadi 2013). This paper focuses on the special context in which variables have a natural ordering, such as when data are collected over time or along a genome. In such a context, it is often reasonable to assume that random variables that are far away in the ordering are less dependent than those that are close together. For example, it is known that genetic mutations that occur close together on a chromosome are more likely to be coinherited than mutations that are located far apart. We propose a method for estimating the precision matrix based on this assumption while also allowing each random variable to have its own notion of closeness.

In general settings where variables do not necessarily have a known order-

ing, two main types of convex methods with strong theoretical results have been developed for introducing sparsity in Ω. The first approach, known as the *graphical lasso* (Yuan & Lin 2007, Banerjee et al. 2008, Friedman et al. 2008, Rothman et al. 2008), performs penalized maximum likelihood, solving $\min_{\Omega>0,\Omega=\Omega^T} \mathcal{L}(\Omega) + \lambda P(\Omega)$, where $\mathcal{L}(\Omega) = -\log\det\Omega + n^{-1}\sum_{i=1}^{n} x_i^T \Omega x_i$ is, up to constants, the negative log-likelihood of a sample of n independent Gaussian random vectors and $P(\Omega)$ is the (vector) ℓ_1-norm of Ω. Zhang & Zou (2014) introduce a new convex loss function called the *D-trace loss* and propose a positive definite precision matrix estimator by minimizing an ℓ_1-penalized version of this loss. The second approach is through penalized pseudo-likelihood, the most well-known of which is called *neighborhood selection* (Meinshausen & Bühlmann 2006). Estimators in this category are usually solved by a column-by-column approach and thus are more amenable to theoretical analysis (Yuan 2010, Cai et al. 2011, Liu & Luo 2012, Liu et al. 2017, Sun & Zhang 2013, Khare et al. 2014). However they are not guaranteed to be positive definite and do not exploit the symmetry of Ω. Peng et al. (2009) propose a partial correlation matrix estimator that develops a symmetric version of neighborhood selection; however, positive definiteness is still not guaranteed.

In the context of variables with a natural ordering, by contrast, almost no work uses convex optimization to flexibly estimate Ω while exploiting the ordering structure. Sparsity is usually induced via the Cholesky decomposition of Σ, which leads to a natural interpretation of sparsity. Consider the Cholesky decomposition $\Sigma = QQ^T$, which implies $\Omega = L^T L$ for $L = Q^{-1}$ for lower triangular matrices Q and L with positive diagonals. The assumption that $X \sim N(0, \Sigma)$ is then equivalent to a set of linear models in terms of rows of L, i.e., $L_{11}X_1 = \varepsilon_1$

and

$$L_{rr}X_r = -\sum_{k=1}^{r-1} L_{rk}X_k + \varepsilon_r \quad r = 2, \ldots, p, \tag{2.1}$$

where $\varepsilon \sim N\left(0, I_p\right)$. Thus, $L_{rk} = 0$ (for $k < r$) can be interpreted as meaning that in predicting X_r from the previous random variables, one does not need to know X_k. This observation has motivated previous work, including Pourahmadi (1999), Wu & Pourahmadi (2003), Huang et al. (2006), Shojaie & Michailidis (2010), Khare et al. (2016). While these methods assume sparsity in L, they do not require local dependence because each variable is allowed to be dependent on predecessors that are distant from it (compare the upper left to the upper right panel of Figure 2.10).

The assumption of "local dependence" can be expressed as saying that each variable X_r can be best explained by exactly its K_r closest predecessors:

$$L_{rr}X_r = -\sum_{k=r-K_r}^{r-1} L_{rk}X_k + \varepsilon_r, \quad \text{for} \quad L_{rk} \neq 0, \quad r - K_r \leq k \leq r-1, \quad r = 2, \ldots, p. \tag{2.2}$$

Note that this does not describe all patterns of a variable depending on its nearby variables. For example, X_r can be dependent on X_{r-2} but not on X_{r-1}. In this case, the dependence is still local, but would not be captured by (2.2). We focus on the restricted class (2.2) since it greatly simplifies the interpretation of the learned dependence structure by capturing the extent of this dependence in a single number K_r, the neighborhood size.

Another desirable property of model (2.2) is that it admits a simple connection between the sparsity pattern of L and the sparsity pattern of the precision matrix Ω in the Gaussian graphical model. In particular, straightforward alge-

bra shows that for $j < k$,

$$L_{kj} = \cdots = L_{pj} = 0 \implies \Omega_{jk} = 0. \tag{2.3}$$

Statistically, this says that if none of the variables $X_k, \ldots, X_p$ depends on X_j in the sense of (2.1), then X_j and X_k are conditionally independent given all other variables.

Bickel & Levina (2008) study theoretical properties in the case that all bandwidths, K_r, are equal, in which case model (2.2) is a K_r-ordered antedependence model (Zimmerman & Nunez-Anton 2009). A banded estimate of L then induces a banded estimate of Ω. The *nested lasso* approach of Levina et al. (2008) provides for "adaptive banding", allowing K_r to vary with r (which corresponds to variable-order antedependence models in Zimmerman & Nunez-Anton 2009); however, the nested lasso is non-convex, meaning that the proposed algorithm does not necessarily minimize the stated objective and theoretical properties of this estimator have not been established.

In this paper, we propose a penalized likelihood approach that provides the flexibility of the nested lasso but is formulated as a convex optimization problem, which allows us to prove strong theoretical properties and to provide an efficient, scalable algorithm for computing the estimator. The theoretical development of our method allows us to make clear comparisons with known results for the graphical lasso (Rothman et al. 2008, Ravikumar et al. 2011) in the non-ordered case. Both methods are convex penalized likelihood approaches, so this comparison highlights the similarities and differences in the ordered and non-ordered problems.

There are two key choices we make that lead to a convex formulation. First, we express the optimization problem in terms of the Cholesky factor L. The

nested lasso and other methods (starting with Pourahmadi 1999) use the modified Cholesky decomposition, $\Omega = T^T D^{-1} T$, where T is a lower-triangular matrix with ones on its diagonal and D is a diagonal matrix with positive entries. While $\mathcal{L}(\Omega)$ is convex in Ω, the negative log-likelihood $\mathcal{L}(T^T D^{-1} T)$ is not jointly convex in T and D. By contrast,

$$\mathcal{L}\left(L^T L\right) = -\log\det\left(L^T L\right) + \frac{1}{n}\sum_{i=1}^{n} x_i^T L^T L x_i = -2\sum_{r=1}^{p}\log L_{rr} + \frac{1}{n}\sum_{i=1}^{n}\|Lx_i\|_2^2 \qquad (2.4)$$

is convex in L. This parametrization is considered in Aragam & Zhou (2015), Khare et al. (2014), and Khare et al. (2016). Maximum likelihood estimation of L preserves the regression interpretation by noting that

$$\mathcal{L}\left(L^T L\right) = -2\sum_{r=1}^{p}\log L_{rr} + \frac{1}{n}\sum_{r=1}^{p}\sum_{i=1}^{n} L_{rr}^2\left(x_{ir} + \sum_{k=1}^{r-1} L_{rk}x_{ik}/L_{rr}\right)^2.$$

This connection has motivated previous work with the modified Cholesky decomposition, in which $T_{rk} = -L_{rk}/L_{rr}$ are the coefficients of a linear model in which X_r is regressed on its predecessors, and $D_{rr} = L_{rr}^{-2}$ corresponds to the error variance. The second key choice is our use of a hierarchical group lasso in place of the nested lasso's nonconvex penalty.

We introduce here some notation used throughout the paper. For two sequences of constants $a(n)$ and $b(n)$, the notation $a(n) = o(b(n))$ means that for every $\varepsilon > 0$, there exists a constant $N > 0$ such that $|a(n)/b(n)| \leq \varepsilon$ for all $n \geq N$. And the notation $a(n) = O(b(n))$ means that there exists a constant $N > 0$ and a constant $M > 0$ such that $|a(n)/b(n)| \leq M$ for all $n \geq N$. For a sequence of random variables $A(n)$, the notation $A(n) = O_P(b(n))$ means that for every $\varepsilon > 0$, there exists a constant $M > 0$ such that $\mathrm{P}(|A(n)/b(n)| > M) \leq \varepsilon$ for all n.

For a vector $v = \left(v_1, \ldots, v_p\right) \in \mathbb{R}^p$, we define $\|v\|_1 = \sum_{j=1}^{p}|v_j|$, $\|v\|_2 = (\sum_{j=1}^{p} v_j^2)^{1/2}$ and $\|v\|_\infty = \max_j |v_j|$. For a matrix $M \in \mathbb{R}^{n\times p}$, we define the element-wise norms

by two vertical bars. Specifically, $\|M\|_\infty = \max_{jk} |M_{jk}|$ and Frobenius norm $\|M\|_F = (\sum_{j,k} M_{jk}^2)^{1/2}$. For $q \geq 1$, we define the matrix-induced (operator) q-norm by three vertical bars: $|||M|||_q = \max_{\|v\|_q=1} \|Mv\|_q$. Important special cases include $|||M|||_2$, also known as the spectral norm, which is the largest singular value of M, as well as $|||M|||_1 = \max_k \sum_{j=1}^p |M_{jk}|$ and $|||M|||_\infty = \max_j \sum_{k=1}^p |M_{jk}|$. Note that $|||M|||_1 = |||M|||_\infty$ when M is symmetric.

Given a p-vector v, a $p \times p$ matrix M, and an index set T, let $v_T = (v_i)_{i \in T}$ be the $|T|$-subvector and M_T the $p \times |T|$ submatrix with columns selected from T. Given a second index set T', let $M_{TT'}$ be the $|T| \times |T'|$ submatrix with rows and columns of M indexed by T and T', respectively. Specifically, we use $L_{r\cdot}$ to denote the r-th row of L.

2.2 Estimator

For a given tuning parameter $\lambda \geq 0$, we define our estimator $\hat{L}$ to be a minimizer of the following penalized negative Gaussian log-likelihood

$$\hat{L} \in \underset{\substack{L: L_{rr}>0 \\ L_{rk}=0 \text{ for } r<k}}{\arg\min} \left\{ -2 \sum_{r=1}^{p} \log L_{rr} + \frac{1}{n} \sum_{i=1}^{n} \|Lx_i\|_2^2 + \lambda \sum_{r=2}^{p} P_r(L_{r\cdot}) \right\}. \tag{2.5}$$

The penalty P_r, which is applied to the r-th row, is defined by

$$P_r(L_{r\cdot}) = \sum_{\ell=1}^{r-1} \left\| W^{(\ell)} * L_{g_{r,\ell}} \right\|_2 = \sum_{\ell=1}^{r-1} \left(\sum_{m=1}^{\ell} w_{\ell m}^2 L_{rm}^2 \right)^{1/2}, \tag{2.6}$$

where $W^{(\ell)} = (w_{\ell 1}, \ldots, w_{\ell \ell}) \in \mathbb{R}^\ell$ is a vector of weights, $*$ denotes element-wise multiplication, and $L_{g_{r,\ell}}$ denotes the vector of elements of L from the group $g_{r,\ell}$, which corresponds to the first ℓ elements in the r-th row (for $1 \leq \ell \leq r-1$):

$$g_{r,\ell} = \{(r, \ell') : \ell' \leq \ell\}.$$

Since $g_{r,1} \subset g_{r,2} \subset \cdots \subset g_{r,r-1}$, each row r of L is penalized with a sum of $r-1$ nested, weighted ℓ_2-norm penalties. This is a hierarchical group lasso penalty (Yuan & Lin 2007, Zhao et al. 2009, Jenatton et al. 2011, Yan & Bien 2015) with group structure conveyed in Figure 2.1.

With $w_{\ell m} > 0$, this nested structure always puts more penalty on those elements that are further away from the diagonal. Since the group lasso has the effect of setting to zero a subset of groups, it is apparent that this choice of groups ensures that whenever the elements in $g_{r,\ell}$ are set to zero, elements in $g_{r,\ell'}$ are also set to zero for all $\ell' \leq \ell$. In other words, for each row of $\hat{L}$, the non-zeros are those elements within some (row-specific) distance of the diagonal. This is in contrast to the ℓ_1-penalty as used in Khare et al. (2016), which produces sparsity patterns with no particular structure (compare the top-left and top-right panels of Figure 2.10).

The choice of weights, $w_{\ell m}$, affects both the empirical and theoretical performance of the estimator. We focus primarily on a quadratically decaying set of weights,

$$w_{\ell m} = \frac{1}{(\ell - m + 1)^2}, \tag{2.7}$$

but also consider the unweighted case (in which $w_{\ell m} = 1$). The decay counteracts the fact that the elements of L appear in differing numbers of groups (for example L_{r1} appears in $r-1$ groups whereas $L_{r,r-1}$ appears in just one group). In a related problem, Bien et al. (2016) choose weights that decay more slowly with $\ell - m$ than (2.7). Our choice makes the enforcement of hierarchy weaker so that our penalty behaves more closely to the lasso penalty (Tibshirani 1996). The choice of weight sequence in (2.7) is more amenable to theoretical analysis; however, in practice the unweighted case is more efficiently implemented and

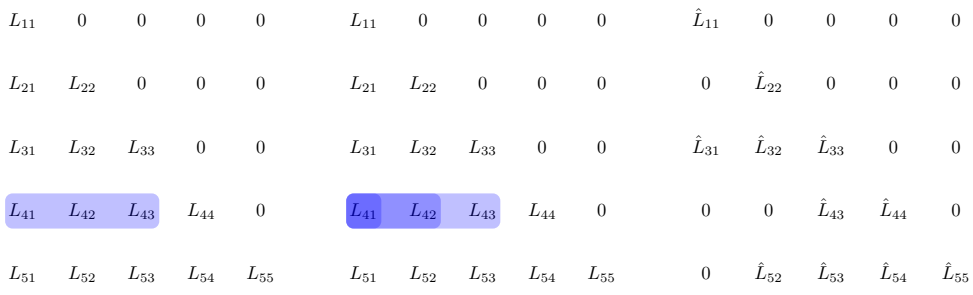

Figure 2.1: There are $\binom{p}{2}$ groups used in the penalty, with each row r having $r-1$ nested groups $g_{r,1} \subset g_{r,2} \subset \cdots \subset g_{r,r-1}$. Left: the group $g_{4,3}$. Middle: the nested group structure $g_{4,1} \subset g_{4,2} \subset g_{4,3}$. Right: A possible sparsity pattern in $\hat{L}$, where elements in $g_{2,1}, g_{4,2}$ (and thus $g_{4,1}$) and $g_{5,1}$ are set to zero.

works well empirically.

Problem (2.5) is convex in L. While $-\log\det(\cdot)$ is strictly convex, $-\sum_r \log(L_{rr})$ is not strictly convex in L. Thus, the arg min in (2.5) may not be unique. In Section 2.4, we provide sufficient conditions to ensure uniqueness with high probability.

In Appendix A.1, we show that (2.5) decouples into p independent subproblems, each of which estimates one row of L. More specifically, let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a sample matrix with independent rows $x_i \sim N(0,\Sigma)$, $\hat{L}_{11} = n^{1/2}(\mathbf{X}_1^T\mathbf{X}_1)^{-1/2}$ and for $r = 2,\ldots,p$,

$$\hat{L}_{r,1:r} = \underset{\beta\in\mathbb{R}^r:\beta_r>0}{\arg\min}\left\{-2\log\beta_r + \frac{1}{n}\|\mathbf{X}_{1:r}\beta\|_2^2 + \lambda\sum_{\ell=1}^{r-1}\left(\sum_{m=1}^{\ell} w_{\ell m}^2\beta_m^2\right)^{1/2}\right\}. \tag{2.8}$$

This observation means that the computation can be easily parallelized, which potentially can achieve a linear speed up with the number of CPU cores. Theoretically, to analyze the properties of $\hat{L}$ it is easier to start by studying an estimator of each row, i.e., a solution to (2.8). We will see in Section 2.4 that problem (2.8) has connections to a penalized regression problem, meaning that both the assumptions and results we can derive are better than if we were working with

a penalty based on Ω.

In light of the regression interpretation of (2.1), $\hat{L}$ provides an interpretable notion of local dependence; however, we can of course also use our estimate of L to estimate Ω: $\hat{\Omega} = \hat{L}^T\hat{L}$. By construction, this estimator is both symmetric and positive definite. Unlike a lasso penalty, which would induce unstructured sparsity in the estimate of L and thus would not be guaranteed to produce a sparse estimate of Ω, the adaptively banded structure in our estimator of L can yield a generally banded $\hat{\Omega}$ with sparsity pattern determined by (2.3) (See the top-left and bottom-left panels in Figure 2.10 for an example).

2.3 Computation

As observed above, we can compute $\hat{L}$ by solving (in parallel across r) problem (2.8). Consider an alternating direction method of multipliers (ADMM) approach that solves the equivalent problem

$$\min_{\beta,\gamma\in\mathbb{R}^r:\beta_r>0}\left\{-2\log\beta_r+\frac{1}{n}\|\mathbf{X}_{1:r}\beta\|_2^2+\lambda\sum_{\ell=1}^{r-1}\left(\sum_{m=1}^{\ell}w_{\ell m}^2\gamma_m^2\right)^{1/2}\quad\text{s.t.}\quad\beta=\gamma\right\}.$$

Algorithm 1 presents the ADMM algorithm, which repeatedly minimizes this problem's augmented Lagrangian over β, then over γ, and then updates the dual variable $u\in\mathbb{R}^r$. The main computational effort in the algorithm is in solving (2.9) and (2.10). Note that (2.9) has a smooth objective function. Straightforward calculus gives the closed-form solution (see Appendix A.2 for detailed derivation),

$$\begin{aligned}\beta_r^{(t+1)}&=\frac{-B-\sqrt{B^2-8A}}{2A}>0\\ \beta_{-r}^{(t+1)}&=-\left(2S_{-r,-r}^{(r)}+\rho I\right)^{-1}\left(2S_{-r,r}^{(r)}\beta_r^{(t+1)}+u_{-r}^{(t)}-\rho\gamma_{-r}^{(t)}\right),\end{aligned}$$

Algorithm 1: ADMM algorithm to solve (2.8)

Require: $\beta^{(0)}$, $\gamma^{(0)}$, $u^{(0)}$, $\rho > 0$, $t = 1$.

1: **repeat**

2:

$$\beta^{(t)} \leftarrow \underset{\beta \in \mathbb{R}^r : \beta_r > 0}{\arg\min} \left\{ -2 \log \beta_r + \frac{1}{n} \|\mathbf{X}_{1:r}\beta\|_2^2 + \left(\beta - \gamma^{(t-1)}\right)^T u^{(t-1)} + \frac{\rho}{2} \left\|\beta - \gamma^{(t-1)}\right\|_2^2 \right\} \tag{2.9}$$

3:

$$\gamma^{(t)} \leftarrow \underset{\gamma \in \mathbb{R}^r}{\arg\min} \left\{ \frac{\rho}{2} \left\|\gamma - \beta^{(t)} - \rho^{-1} u^{(t-1)}\right\|_2^2 + \lambda \sum_{\ell=1}^{r-1} \left(\sum_{m=1}^{\ell} w_{\ell m}^2 \gamma_m^2 \right)^{1/2} \right\} \tag{2.10}$$

4: $u^{(t)} \leftarrow u^{(t-1)} + \rho\left(\beta^{(t)} - \gamma^{(t)}\right)$

5: $t \leftarrow t + 1$

6: **until** convergence

7: **return** $\gamma^{(t)}$

where

$$\begin{aligned} S^{(r)} &= \frac{1}{n} \mathbf{X}_{1:r}^T \mathbf{X}_{1:r} \\ A &= 4S_{r,-r}^{(r)} \left(2S_{-r,-r}^{(r)} + \rho I\right)^{-1} S_{-r,r}^{(r)} - 2S_{r,r}^{(r)} - \rho < 0 \\ B &= 2S_{r,-r}^{(r)} \left(2S_{-r,-r}^{(r)} + \rho I\right)^{-1} \left(u_{-r}^{(t)} - \rho \gamma_{-r}^{(t)}\right) - u_r^{(t)} + \rho \gamma_r^{(t)}. \end{aligned}$$

The closed-form update above involves matrix inversion. With $\rho > 0$, the matrix $2S_{-r,-r}^{(r)} + \rho I$ is invertible even when $r > n$. Since determining a good choice for the ADMM parameter ρ is in general difficult, we adapt the dynamic ρ updating scheme described in Section 3.4.1 of Boyd et al. (2011).

Solving (2.10) requires evaluating the proximal operator of the hierarchical

group lasso with general weights. We adopt the strategy developed in Bien et al. (2016) (based on a result of Jenatton et al. 2011), which solves the dual problem of (2.10) by performing Newton's method on at most $r-1$ univariate functions. The detailed implementation is given in Algorithm 5 in Appendix A.3. Each application of Newton's method corresponds to performing an elliptical projection, which is a step of blockwise coordinate ascent on the dual of (2.10) (see Appendix A.4 for details). Finally we observe in Algorithm 2 that for the unweighted case ($w_{\ell m} = 1$), solving (2.10) is remarkably efficient.

Algorithm 2: Algorithm for solving (2.10) for unweighted estimator

---

Require: $\beta^{(t)}, u^{(t-1)} \in \mathbb{R}^r$, $\lambda, \rho > 0$.

1: Initialize $\gamma^{(t)} = \beta^{(t)} + u^{(t-1)}/\rho$ and $\tau = \lambda/\rho$

2: **for** $\ell = 1, \ldots, r-1$ **do**

$$\left(\gamma^{(t)}\right)_{1:\ell} \leftarrow \left(1 - \frac{\tau}{\left\|\left(\gamma^{(t)}\right)_{1:\ell}\right\|_2}\right)_+ \left(\gamma^{(t)}\right)_{1:\ell}$$

3: **return** $\gamma^{(t)}$.

The `R` package `varband` provides `C++` implementations of Algorithms 1 and 2.

2.4 Statistical properties

In this section we study the statistical properties of our estimator. In what follows, we consider a lower triangular matrix L having row-specific bandwidths, K_r. The first $J_r = r - 1 - K_r$ elements of row r are zero, and the band of non-

zero off-diagonals (of size K_r) is denoted $\mathcal{I}_r = \{J_r + 1, \ldots, r-1\}$. We also denote $\mathcal{I}_r^c = \{1, 2, \ldots, r\} \setminus \mathcal{I}_r$. See Figure 2.2 for a graphical example of $K_5, J_5, \mathcal{I}_4$, and $\mathcal{I}_4^c$.

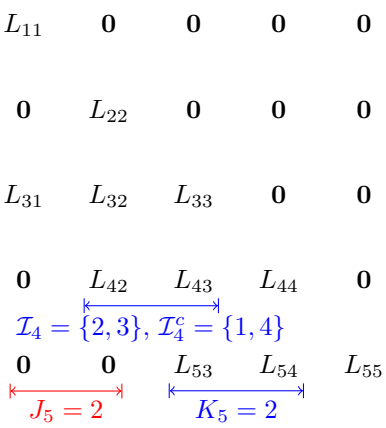

Figure 2.2: Schematic showing $J_r, K_r, \mathcal{I}_r$, and $\mathcal{I}_r^c$.

Our theoretical analysis is built on the following assumptions:

A1 *Gaussian assumption*: The sample matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has n independent rows with each row x_i drawn from $N(\mathbf{0}, \Sigma)$.

A2 *Sparsity assumption*: The true Cholesky factor $L \in \mathbb{R}^{p \times p}$ is the lower triangular matrix with positive diagonal elements such that the precision matrix $\Omega = \Sigma^{-1} = L^T L$. The matrix L has row-specific bandwidths K_r such that $L_{rj} = 0$ for $0 < j < r - K_r$.

A3 *Irrepresentable condition*: There exists some $\alpha \in (0, 1]$ such that

$$\max_{2 \le r \le p} \max_{\ell \in \mathcal{I}_r^c} \left\| \Sigma_{\ell \mathcal{I}_r} \left(\Sigma_{\mathcal{I}_r \mathcal{I}_r} \right)^{-1} \right\|_1 \le \frac{6}{\pi^2} (1 - \alpha)$$

A4 *Bounded singular values*: There exists a constant κ such that

$$0 < \kappa^{-1} \le \sigma_{\min}(L) \le \sigma_{\max}(L) \le \kappa$$

When $\max_r K_r < n$, the Gaussianity assumption **A1** implies that $\mathbf{X}_{\mathcal{I}_r}$ has full column rank for all r with probability one. Our analysis applies to the general high-dimensional scaling scheme where $K_r = K_r(n)$ and $p = p(n)$ can grow with n.

For $r = 2, \ldots, p$ and $\ell \in \mathcal{I}_r^c = \{1, \ldots, J_r, r\}$, let

$$\theta_r^{(\ell)} := \mathrm{Var}\left(X_\ell | X_{\mathcal{I}_r}\right) \quad \text{and} \quad \theta_r := \max_{\ell \in \mathcal{I}_r^c} \theta_r^{(\ell)}.$$

By Assumption **A1**, $\theta_r^{(\ell)} = \Sigma_{\ell\ell} - \Sigma_{\ell \mathcal{I}_r} \left(\Sigma_{\mathcal{I}_r \mathcal{I}_r}\right)^{-1} \Sigma_{\mathcal{I}_r \ell}$ represents the noise variance when regressing X_ℓ on $X_{\mathcal{I}_r}$, i.e., for $\ell = 1, \ldots, J_r, r$,

$$X_\ell = \Sigma_{\ell \mathcal{I}_r} (\Sigma_{\mathcal{I}_r \mathcal{I}_r})^{-1} X_{\mathcal{I}_r}^T + E_\ell \qquad \text{with} \qquad E_\ell \sim N\left(0, \theta_r^{(\ell)}\right). \tag{2.11}$$

In words, $\theta_r^{(\ell)}$ measures the degree to which X_ℓ cannot be explained by the variables in the support and θ_r is the maximum such value over all ℓ outside of the support $\mathcal{I}_r$ in the r-th row. Intuitively, the difficulty of the estimation problem increases with θ_r. Note that for $r = 1, \ldots, p$, (2.1) implies $\theta_r^{(r)} = 1/L_{rr}^2$.

Assumption **A3** (along with the $\beta_{\min}$ condition) is essentially a necessary and sufficient condition for support recovery of lasso-type methods (see, e.g., Zhao & Yu 2006, Meinshausen & Bühlmann 2006, Wainwright 2009, Van de Geer & Bühlmann 2009, Ravikumar et al. 2011). The constant $\alpha \in (0, 1]$ is usually referred to as the irrepresentable (incoherence) constant (Wainwright 2009). Intuitively, the irrepresentable condition requires low correlations between signal and noise predictors, and thus a value of α that is close to 1 implies that recovering the support is easier to achieve. The constant $6\pi^{-2}$ is determined by the choice of weight (2.7) and can be eliminated by absorbing its reciprocal into the definition of the weights $w_{\ell m}$. Doing so, one finds that our irrepresentable condition is essentially the same as the one found in the regression setting (Wainwright 2009) despite the fact that our goal is estimating a precision matrix.

Assumption **A4** is a bounded singular value condition. Recalling that $\Omega = L^T L$,

$$0 < \kappa^{-2} \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq \kappa^2, \tag{2.12}$$

which is equivalent to the commonly used bounded eigenvalue condition in other literatures.

2.4.1 Row-specific results

We start by analyzing support recovery properties of our estimator for each row, i.e., the solution to the subproblem (2.8). For $r > n$, the Hessian of the negative log-likelihood is not positive definite, meaning that the objective function may not be strictly convex in β and the solution not necessarily unique. Intuitively, if the tuning parameter λ is large, the resulting row estimate $\hat{L}_{r\cdot}$ is sparse and thus includes most variation in a small subset of the r variables. More specifically, for large λ, $\hat{\mathcal{I}}_r \subseteq \mathcal{I}_r$ and thus by Assumption **A1**, $\mathbf{X}_{\hat{\mathcal{I}}_r}$ has full rank, which implies that $\hat{L}_{r\cdot}$ is unique. The series of technical lemmas in Appendix A.5 precisely characterizes the solution.

The first part of the theorem below shows that with an appropriately chosen tuning parameter λ the solution to (2.8) is sparse enough to be unique and that we will not over-estimate the true bandwidth. Knowing that the support of the unique row estimator $\hat{L}_{r\cdot}$ is contained in the true support reduces the dimension of the parameter space, and thus leads to a reasonable error bound. Of course, if our goal were simply to establish the uniqueness of $\hat{L}_{r\cdot}$ and that $\hat{K}_r \leq K_r$, we could trivially take $\lambda = \infty$ (resulting in $\hat{K}_r = 0$). The latter part of the theorem thus goes on to provide a choice of λ that is sufficiently small to guarantee that

$\hat{K}_r = K_r$ (and, furthermore, that the signs of all non-zeros are correctly recovered).

Theorem 1. *Consider the family of tuning parameters*

$$\lambda = \frac{8}{\alpha}\sqrt{\frac{\theta_r \log r}{n}} \tag{2.13}$$

and weights given by (2.7)*. Under Assumptions **A1**–**A4**, if the tuple* (n, J_r, K_r) *satisfies*

$$n > \alpha^{-2}\left(3\pi^2 K_r + 8\right)\theta_r \kappa^2 \log J_r, \tag{2.14}$$

then with probability greater than $1 - c_1 \exp\{-c_2 \min(K_r, \log J_r)\} - 7\exp(-c_3 n)$ *for some constants* c_1, c_2, c_3 *independent of* n *and* J_r*, the following properties hold:*

1. *The row problem* (2.8) *has a unique solution* $\hat{L}_{r\cdot}$ *and* $\hat{K}_r \le K_r$.
2. *The estimate* $\hat{L}_{r\cdot}$ *satisfies the element-wise* ℓ_∞ *bound,*

$$\left\|\hat{L}_{r\cdot} - L_{r\cdot}\right\|_\infty \le \lambda\left(4\left|\!\left|\!\left|(\Sigma_{I_r I_r})^{-1}\right|\!\right|\!\right|_\infty + 5\kappa^2\right). \tag{2.15}$$

3. *If in addition,*

$$\min_{j \ge J_r + 1}\left|L_{rj}\right| > \lambda\left(4\left|\!\left|\!\left|(\Sigma_{I_r I_r})^{-1}\right|\!\right|\!\right|_\infty + 5\kappa^2\right), \tag{2.16}$$

then exact signed support recovery holds: For all $j \le r$, $\operatorname{sign}(\hat{L}_{rj}) = \operatorname{sign}(L_{rj})$.

Proof. See Appendix A.6. □

In the classical setting where the ambient dimension r is fixed and the sample size n is allowed to go to infinity, $\lambda \to 0$ and the above scaling requirement is satisfied. By (2.15) the row estimator $\hat{L}_{r\cdot}$ is consistent as is the classical maximum likelihood estimator. Moreover, it recovers the true support since (2.16) holds automatically. In high-dimensional scaling, however, both n and r are allowed

to change, and we are interested in the case where r can grow much faster than n. Theorem 1 shows that, if $\left\|(\Sigma_{I_r I_r})^{-1}\right\|_\infty = O(1)$ and if n can grow as fast as $K_r \log J_r$, then the row estimator $\hat{L}_{r\cdot}$ still recovers the exact support of $L_{r\cdot}$ when the signal is at least $O(\sqrt{\frac{\log r}{n}})$ in size, and the estimation error $\max_j |\hat{L}_{rj} - L_{rj}|$ is $O(\sqrt{\frac{\log r}{n}})$. Intuitively, for the row estimator to detect the true support, we require that the true signal be sufficiently large. The condition (2.16) imposes limitations on how fast the signal is allowed to decay, which is the analogue to the commonly known "$\beta_{\min}$ condition" that is assumed for establishing support recovery of the lasso.

Remark 2. *Both the choice of tuning parameter* (2.13) *and the error bound* (2.15) *depend on the true covariance matrix via* θ_r*. This quantity can be bounded by* κ^2 *as in* (2.12) *using the fact that* $(\Sigma_{I_r I_r})^{-1}$ *is positive definite:*

$$\theta_r = \max_{\ell \in I_r^c} \theta_r^{(\ell)} = \max_{\ell \in I_r^c} \left\{ \Sigma_{\ell\ell} - \Sigma_{\ell I_r} \left(\Sigma_{I_r I_r}\right)^{-1} \Sigma_{I_r \ell} \right\} \leq \max_{\ell \in I_r^c} \Sigma_{\ell\ell} \leq \kappa^2.$$

The proof of Theorem 1 shows that the results in this theorem still hold true if we replace θ_r *by* κ^2*. This observation leads to the fact that we can select a tuning parameter having the properties of the theorem that does not depend on the unknown sparsity level* K_r*. Therefore, our estimator is adaptive to the underlying unknown bandwidths.*

Connections to the regression setting

In (2.1) we showed that estimation of the r-th row of L can be interpreted as a regression of X_r on its predecessors. It is thus very interesting to compare Theorem 1 to the standard high-dimensional regression results. Consider the following linear model of a vector $\mathbf{y} \in \mathbb{R}^n$ of the form

$$\mathbf{y} = \mathbf{Z}\eta + \omega \qquad \omega \sim N(\mathbf{0}, \sigma^2 I_n) \tag{2.17}$$

where $\eta \in \mathbb{R}^p$ is the unknown but fixed parameter to estimate, $\mathbf{Z} \in \mathbb{R}^{n \times p}$ is the design matrix with each row an observation of p predictors, σ^2 is the variance of the zero-mean additive noise ω. A standard approach in the high-dimensional setting where $p \gg n$ is the lasso (Tibshirani 1996), which solves the convex optimization problem,

$$\min_{\eta \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{Z}\eta\|_2^2 + \lambda \|\eta\|_1 , \tag{2.18}$$

where $\lambda > 0$ is a regularization parameter. In the setting where η is assumed to be sparse, the lasso solution is known to be able to successfully recover the signed support of the true η with high probability when λ is of the scale $\sigma \sqrt{\frac{\log p}{n}}$ and certain technical conditions are satisfied (Wainwright 2009).

Despite the added complications of working with the log term in the objective of (2.8), Theorem 1 gives a clear indication that, in terms of difficulty of support recovery, the row estimate problem (2.8) is essentially the same as a lasso problem with random design, i.e., with each row $z_i \sim N(\mathbf{0}, \Sigma)$ (Theorem 3, Wainwright 2009). Indeed, a comparison shows that the two irrepresentable conditions are equivalent. Moreover, θ_r plays the same role as Wainwright (2009)'s $\max_i \left(\Sigma_{S^c S^c} - \Sigma_{S^c S} \left(\Sigma_{SS} \right)^{-1} \Sigma_{S S^c} \right)_{ii}$, a threshold constant of the conditional covariance, where S is the support of the true η.

Städler et al. (2010) introduce an alternative approach to the lasso, in the context of penalized mixture regression models, that solves the optimization problem,

$$(\hat{\phi}, \hat{\rho}) = \arg\min_{\phi, \rho} \left\{ -2 \log \rho + \frac{1}{n} \|\rho \mathbf{y} + \mathbf{Z}\phi\|_2^2 + \lambda \|\phi\|_1 \right\}, \tag{2.19}$$

where $\hat{\sigma} = \hat{\rho}^{-1}$ and $\hat{\eta} = -\hat{\phi}/\hat{\rho}$. Note that (2.19) basically coincides with (2.8) except for the penalty.

In Städler et al. (2010), the authors study the asymptotic and non-asymptotic properties of the ℓ_1-penalized estimator for the general mixture regression models where the loss functions are non-convex. The theoretical properties of (2.19) are studied in Sun & Zhang (2010), which partly motivates the scaled lasso (Sun & Zhang 2012).

The theoretical work of Sun & Zhang (2010) differs from ours both in that they study the ℓ_1 penalty (instead of the hierarchical group lasso) and in their assumptions. The nature of our problem requires the sample matrix to be random (as in **A1**), while Sun & Zhang (2010) considers the fixed design setting, which does not apply in our context. Moreover, they provide prediction consistency and a deviation bound of the regression parameters estimation in ℓ_1 norm. We give exact signed support recovery results for the regression parameters as well as estimation deviation bounds in various norm criteria. Also, they take an asymptotic point of view while we give finite sample results.

2.4.2 Matrix bandwidth recovery result

With the properties of the row estimators in place, we are ready to state results about estimation of the matrix L. The following theorem gives an analogue to Theorem 1 in the matrix setting. Under similar conditions, with one particular choice of tuning parameter, the estimator recovers the true bandwidth for all rows adaptively with high probability.

Theorem 3. *Let* $\theta = \max_r \theta_r$ *and* $K = \max_r K_r$*, and take*

$$\lambda = \frac{8}{\alpha}\sqrt{\frac{2\theta \log p}{n}} \tag{2.20}$$

and weights given by (2.7)*. Under Assumptions* ***A1****–****A4****, if* (n, p, K) *satisfies*

$$n > \alpha^{-2}\theta\kappa^2\left(12\pi^2 K + 32\right)\log p, \tag{2.21}$$

then with probability greater than $1 - cp^{-1}$ *for some constant* c *independent of* n *and* p*, the following properties hold:*

1. *The estimator* $\hat{L}$ *is unique, and it is at least as sparse as* L*, i.e.,* $\hat{K}_r \le K_r$ *for all* r*.*
2. *The estimator* $\hat{L}$ *satisfies the element-wise* ℓ_∞ *bound,*

$$\left\|\hat{L} - L\right\|_\infty \le \lambda\left(4\max_r \left\|\left(\Sigma_{\mathcal{I}_r\mathcal{I}_r}\right)^{-1}\right\|_\infty + 5\kappa^2\right). \tag{2.22}$$

3. *If in addition,*

$$\min_r \min_{j \ge J_r + 1} \left|L_{rj}\right| > \lambda\left(4\max_r \left\|\left(\Sigma_{\mathcal{I}_r\mathcal{I}_r}\right)^{-1}\right\|_\infty + 5\kappa^2\right), \tag{2.23}$$

then exact signed support recovery holds: $\operatorname{sign}(\hat{L}_{rj}) = \operatorname{sign}(L_{rj})$ *for all* r *and* j*.*

Proof. See Appendix A.7. □

As discussed in Remark 2, we can replace θ with its upper bound κ^2, and the results remain true. This theorem shows that one can properly estimate the sparsity pattern across all rows exactly using only one tuning parameter chosen without any prior knowledge of the true bandwidths. In Section 2.4.1, we noted that the conditions required for support recovery and the element-wise ℓ_∞ error bound for estimating a row of L is similar to those of the lasso in the regression setting. A union bound argument allows us to translate this into exact bandwidth recovery in the matrix setting and to derive a reasonable convergence rate under conditions as mild as that of a lasso problem with random design.

This technique is similar in spirit to neighborhood selection (Meinshausen & Bühlmann 2006), though our approach is likelihood-based.

Comparing (2.21) to (2.14), we see that the sample size requirement for recovering L is determined by the least sparse row. While intuitively one would expect the matrix problem to be harder than any single row problem, we see that in fact the two problems are basically of the same difficulty (up to a multiplicative constant).

In the setting where variables exhibit a natural ordering, Shojaie & Michailidis (2010) proposed a penalized likelihood framework like ours to estimate the structure of directed acyclic graphs (DAGs). Their method focuses on variables which are standardized to have unit variance. In this special case, penalized likelihood does not involve the log-determinant term and under similar assumptions to ours, they proved support recovery consistency. However, they use lasso and adaptive lasso (Zou 2006) penalties, which do not have the built-in notion of local dependence. Since these ℓ_1-type penalties do not induce structured sparsity in the Cholesky factor, the resulting precision matrix estimate is not necessarily sparse. By contrast, our method does not assume unit variances and learns an adaptively banded structure for $\hat{L}$ that leads to a sparse $\hat{\Omega}$ (thereby encoding conditional dependencies).

To study the difference between the ordered and non-ordered problems, we compare our method with Ravikumar et al. (2011), who studied the graphical lasso estimator in a general setting where variables are not necessarily ordered. Let $\mathcal{S}$ index the edges of the graph specified by the sparsity pattern of $\Omega = \Sigma^{-1}$. The sparsity recovery result and convergence rate are established under an

irrepresentable condition imposed on $\Gamma = \Sigma \otimes \Sigma \in \mathbb{R}^{p^2 \times p^2}$:

$$\max_{e \in \mathcal{S}^c} \left\| \Gamma_{e\mathcal{S}} \left(\Gamma_{\mathcal{S}\mathcal{S}} \right)^{-1} \right\|_1 \leq (1 - \alpha) \tag{2.24}$$

for some $\alpha \in (0, 1]$. Our Assumption **A3** is on each variable through the entries of the true covariance Σ while (2.24) imposes such a condition on the edge variables $Y_{(j,k)} = X_j X_k - \mathrm{E}\left(X_j X_k\right)$, resulting in a vector ℓ_1-norm restriction on a much larger matrix Γ, which can be more restrictive for large p. More specifically, condition (2.24) arises in Ravikumar et al. (2011) to tackle the analysis of the $\log\det\Omega$ term in the graphical lasso problem. By contrast, in our setting the parameterization in terms of L means that the $\log\det$ term is simply a sum of log terms on diagonal elements and is thus easier to deal with, leading to the milder irrepresentable assumption. Another difference is that they require the sample size $n > c\kappa_\Gamma^2 d^2 \log p$ for some constant c. The quantity d measures the maximum number of non-zero elements in each row of the true Σ, which in our case is $2K + 1$, and $\kappa_\Gamma = \left\| \left(\Gamma_{\mathcal{S}\mathcal{S}} \right)^{-1} \right\|_\infty$ can be much larger than κ^2. Thus, comparing to (2.21), one finds that their sample size requirement is much more restrictive. A similar comparison could also be made with the lasso penalized D-trace estimator (Zhang & Zou 2014), whose irrepresentable condition involves $\Gamma = (\Sigma \otimes I + I \otimes \Sigma)/2 \in \mathbb{R}^{p^2 \times p^2}$. Of course, the results in both Ravikumar et al. (2011) and Zhang & Zou (2014) apply to estimators invariant to permutation of variables; additionally, the random vector only needs to satisfy an exponential-type tail condition.

2.4.3 Precision matrix estimation consistency

Although our primary target of interest is L, the parameterization $\Omega = L^T L$ makes it natural for us to try to connect our results of estimating L with the

vast literature in directly estimating Ω, which is the standard estimation target when the known ordering is not available. In this section, we consider the estimation consistency of Ω using the results we obtained for L. The following theorem gives results of how well $\hat{\Omega} = \hat{L}^T \hat{L}$ performs in estimating the true precision matrix $\Omega = L^T L$ in terms of various matrix norm criteria.

Theorem 4. *Let $\theta = \max_r \theta_r$, $K = \max_r K_r$ and $s = \sum_r K_r$ denote the total number of non-zero off-diagonal elements in L. Define $\zeta_\Sigma = \frac{8\sqrt{2\theta}}{\alpha}\left(4 \max_r \left\|(\Sigma_{I_r I_r})^{-1}\right\|_\infty + 5\kappa^2\right)$. Under the assumptions in Theorem 3, the following deviation bounds hold with probability greater than $1 - cp^{-1}$ for some constant c independent of n and p:*

$$
\begin{aligned}
\left\|\hat{\Omega} - \Omega\right\|_\infty &\leq 2\zeta_\Sigma \|L\|_\infty \sqrt{\frac{\log p}{n}} + \zeta_\Sigma^2 (K+1) \frac{\log p}{n}, \\
\left\|\hat{\Omega} - \Omega\right\|_\infty &\leq 2\zeta_\Sigma \|L\|_\infty (K+1) \sqrt{\frac{\log p}{n}} + \zeta_\Sigma^2 (K+1)^2 \frac{\log p}{n}, \\
\left\|\hat{\Omega} - \Omega\right\|_2 &\leq 2\zeta_\Sigma \|L\|_\infty (K+1) \sqrt{\frac{\log p}{n}} + \zeta_\Sigma^2 (K+1)^2 \frac{\log p}{n}, \\
\left\|\hat{\Omega} - \Omega\right\|_F &\leq 2\kappa\zeta_\Sigma \sqrt{\frac{(s+p)\log p}{n}} + \zeta_\Sigma^2 (K+1) \sqrt{s+p} \frac{\log p}{n}.
\end{aligned}
$$

When the quantities ζ_Γ, $\|L\|_\infty$, and κ are treated as constants, these bounds can be summarized more succinctly as follows:

Proof. See Appendix A.8. □

Corollary 5. *Using the notation and conditions in Theorem 4, if ζ_Γ, $\|L\|_\infty$, and κ remain constant, then the scaling $(K+1)^2 \log p = o(n)$ is sufficient to guarantee the following*

estimation error bounds:

$$\begin{aligned}
\left\|\hat{\Omega} - \Omega\right\|_{\infty} &= O_P\left(\sqrt{\frac{\log p}{n}}\right), \\
\left|\!\left|\!\left|\hat{\Omega} - \Omega\right|\!\right|\!\right|_{\infty} &= O_P\left((K+1)\sqrt{\frac{\log p}{n}}\right), \\
\left|\!\left|\!\left|\hat{\Omega} - \Omega\right|\!\right|\!\right|_{2} &= O_P\left((K+1)\sqrt{\frac{\log p}{n}}\right), \\
\left\|\hat{\Omega} - \Omega\right\|_{F} &= O_P\left(\sqrt{\frac{(s+p)\log p}{n}}\right).
\end{aligned}$$

The conditions for these deviation bounds to hold are those required for support recovery as in Theorem 3. In many cases where estimation consistency is more of interest than support recovery, we can still deliver the desired error rate in Frobenius norm, matching the rate derived in Rothman et al. (2008). In particular, we can drop the strong irrepresentable assumption (**A3**) and weaken the Gaussian assumption (**A1**) to the following marginal sub-Gaussian assumption:

A5 *Marginal sub-Gaussian assumption*: The sample matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has n independent rows with each row drawn from the distribution of a zero-mean random vector $X = (X_1, \cdots, X_p)^T$ with covariance Σ and sub-Gaussian marginals, i.e.,

$$\mathrm{E} \exp\left(tX_j / \sqrt{\Sigma_{jj}}\right) \leq \exp\left(Ct^2\right)$$

for all $j = 1, \ldots, p$, $t \geq 0$ and for some constant $C > 0$ that does not depend on j.

Theorem 6. *Under Assumption **A2**, **A4** and **A5**, with tuning parameter λ of scale $\sqrt{\frac{\log p}{n}}$ and weights as in* (2.7)*, the scaling $(s+p)\log p = o(n)$ is sufficient for the fol-*

lowing estimation error bounds in Frobenius norm to hold:

$$\left\|\hat{L} - L\right\|_F = O_P\left(\sqrt{\frac{(s+p)\log p}{n}}\right),$$
$$\left\|\hat{\Omega} - \Omega\right\|_F = O_P\left(\sqrt{\frac{(s+p)\log p}{n}}\right).$$

Proof. See Appendix A.9. □

The rates in Corollary 5 (and Theorem 6) essentially match the rates obtained in methods that directly estimate Ω (e.g., the graphical lasso estimator, studied in Rothman et al. 2008, Ravikumar et al. 2011, and the column-by-column methods as in Cai et al. 2011, Liu et al. 2017, and Sun & Zhang 2013). However, the exact comparison in rates with these methods is not straightforward. First, the targets of interest are different. In the setting where the variables have a known ordering, we are more interested in the structural information among variables that is expressed in L, and thus accurate estimation of L is more important. When such ordering is not available as considered in Rothman et al. (2008), Cai et al. (2011), Liu et al. (2017) and so on, however, the conditional dependence structure encoded by the sparsity pattern in Ω is more of interest, and the accuracy of directly estimating Ω is the focus. Moreover, deviation bounds of different methods are built upon assumptions that treat different quantities as constants. Quantities that are assumed to remain constant in the analysis of one method might actually be allowed to scale with ambient dimension in a non-trivial manner in another method, which makes direct rate comparison among different methods complicated and less illuminating.

Our analysis can be extended to the unweighted version of our estimator, i.e., with weight $w_{\ell m} = 1$, but under more restrictive conditions and with slower rates of convergence. Specifically, Assumption **A3** becomes $\max_{\ell \in \mathcal{I}_r^c} \left\|\Sigma_{\ell \mathcal{I}_r} \left(\Sigma_{\mathcal{I}_r \mathcal{I}_r}\right)^{-1}\right\|_1 \leq$

$(1-\alpha)/K_r$ for each $r = 2, \ldots, p$. With the same tuning parameter choice (2.13) and (2.20), the terms of K_r and K in sample size requirements (2.14) and (2.21) are replaced with K_r^2 and K^2, respectively. The estimation error bounds in all norms are multiplied by an extra factor of K. All of the above indicates that in highly sparse situations (in which K is very small), the unweighted estimator has very similar theoretical performance to the weighted estimator.

2.5 Simulation study

In this section we study the empirical performance of our estimators (both with weights as in (2.7) and with no weights, i.e., $w_{\ell m} = 1$) on simulated data. For comparison, we include two other sparse precision matrix estimators designed for the ordered-variable case:

- **Non-Adaptive Banding** (Bickel & Levina 2008): This method estimates L as a lower-triangular matrix with a fixed bandwidth K applying across all rows. The regularization parameter used in this method is the fixed bandwidth K.
- **Nested Lasso** (Levina et al. 2008): This method yields an adaptive banded structure by solving a set of penalized least-squares problems (both the loss function and the nested-lasso penalty are non-convex). The regularization parameter controls the amount of penalty and thus the sparsity level of the resulting estimate.

All simulations are run at a sample size of $n = 100$, where each sample is drawn independently from the p-dimensional normal distribution $N(\mathbf{0}, (L^T L)^{-1})$.

We compare the performance of our estimators with the methods above both in terms of support recovery (in Section 2.5.1) and in terms of how well $\hat{L}$ estimates L (in Section 2.5.2). For support recovery, we consider $p = 200$ and for estimation accuracy, we consider $p = 50, 100, 200$, which corresponds to settings where $p < n$, $p = n$, and $p > n$, respectively.

We simulate under the following models for L. We adapt the parameterization $L = D^{-1}T$ as in Khare et al. (2016), where D is a diagonal matrix with diagonal elements drawn randomly from a uniform distribution on the interval $[2, 5]$, and T is a lower-triangular matrix with ones on its diagonal and off-diagonal elements defined as follows:

- **Model 1**: Model 1 is at one extreme of bandedness of the Cholesky factor L, in which we take the lower triangular matrix $L \in \mathbb{R}^{p \times p}$ to have a strictly banded structure, with each row having the same bandwidth $K_r = K = 1$ for all r. Specifically, we take $T_{r,r} = 1$, $T_{r,r-1} = 0.8$ and $T_{r,j} = 0$ for $j < r - 1$.
- **Model 2**: Model 2 is at the other extreme, in which we allow K_r to vary with r. We take T to be a block diagonal matrix with 5 blocks, each of size $p/5$. Within each block, with probability 0.5 each row r is assigned with a non-zero bandwidth that is randomly drawn from a uniform distribution on $\{1, \ldots, r-1\}$ (for $r > 1$). Each non-zero element in T is then drawn independently from a uniform distribution on the interval $[0.1, 0.4]$, and is assigned with a positive/negative sign with probability 0.5.
- **Model 3**: Model 3 is a denser and thus more challenging version of Model 2, with T a block diagonal matrix with only 2 blocks. Each of the blocks is of size $p/2$ but is otherwise generated as in Model 2.
- **Model 4**: Model 4 is a dense block diagonal model. The matrix T has a

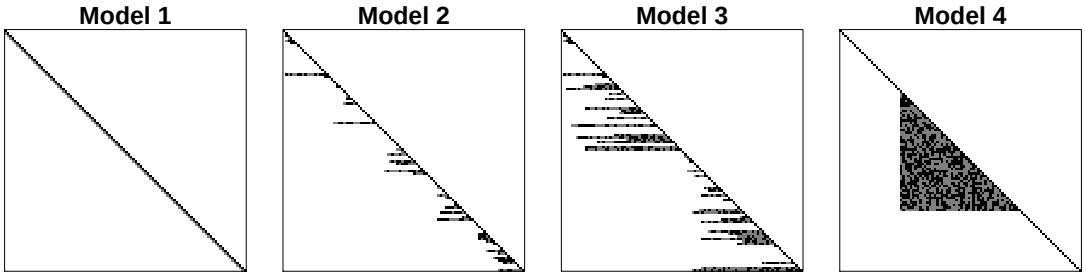

Figure 2.3: Schematic of four simulation scenarios with $p = 100$: (from left to right) Model 1 is strictly banded, Model 2 has small variable bandwidth, Model 3 has large variable bandwidth, and Model 4 is block-diagonal. Black, gray, and white stand for positive, negative, and zero entries, respectively. The proportion of elements that are non-zero is 4%, 6%, 15%, and 26%, respectively.

completely dense lower-triangular block from the $p/4$-th row to the $3p/4$-th row and is zero everywhere else. Within this block, all off-diagonal elements are drawn uniformly from $[0.1, 0.2]$, and positive/negative signs are then assigned with probability 0.5.

Model 1 is a stationary autoregressive model of order 1. By the regression interpretation (2.1), for each r, it can be verified that the autoregressive polynomial of the r-th row of Models 2, 3, and 4 has all roots outside the unit circle, which characterizes stationary autoregressive models of orders equal to the corresponding row-wise bandwidths. See Figure 2.3 for examples of the four sparsity patterns for $p = 100$. The non-adaptive banding method should benefit from Model 1 while the nested lasso and our estimators are expected to perform better in the other three models where each row has its own bandwidth.

For all four models and every value of p considered, we verified that Assumptions **A3** and **A4** hold and then simulated $n = 100$ observations according to each of the four models based on Assumption **A1**.

2.5.1 Support recovery

We first study how well the different estimators identify zeros in the four models above. We generate $n = 100$ random samples from each model with $p = 200$. The tuning parameter $\lambda \geq 0$ in (2.5) measures the amount of regularization and determines the sparsity level of the estimator. We use 100 tuning parameter values for each estimator and repeat the simulation 10 times.

Figure 2.4 shows the sensitivity (fraction of true non-zeros that are correctly recovered) and specificity (fraction of true zeros that are correctly set to zero) of each method parameterized by its tuning parameter (in the case of non-adaptive banding, the parameter is the bandwidth itself, ranging from 0 to $p-1$). Each set of 10 curves of the same color corresponds to the results of one estimator, and each curve within the set corresponds to the result of one draw from 10 simulations. Curves closer to the upper-right corner indicate better classification performance (the $x + y = 1$ line corresponds to random guessing).

The sparsity level of the non-adaptive banding estimator depends only on the pre-specified bandwidth (which is the method's tuning parameter) and not on the data itself. Consequently, the sensitivity-specificity curves for the non-adaptive banding do not vary across replications when simulating from a particular underlying model. The sparsity levels of the nested lasso and our methods, by contrast, hinge on the data, thus giving a different curve for each replication.

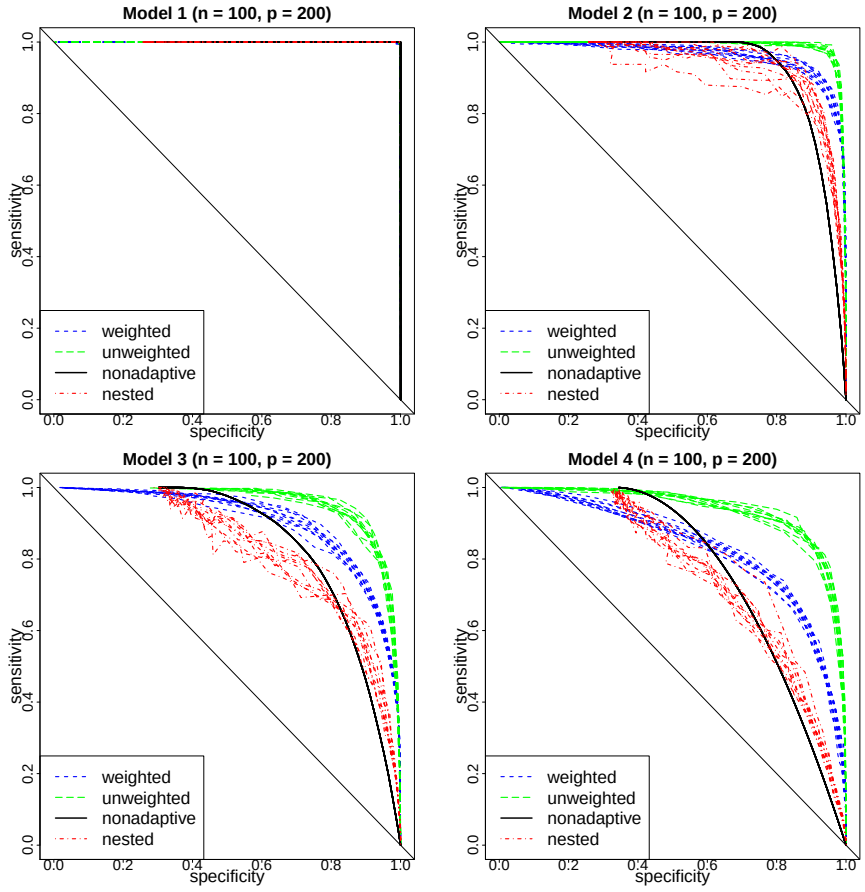

Figure 2.4: ROC curves showing support recovery when the true L (top-left) is strictly banded, (top-right) has small variable bandwidth, (bottom-left) has large variable bandwidth, and (bottom-right) is block-diagonal, over 10 replications.

In practice, we find that our methods and the nested lasso sometimes produce entries with very small, but non-zero, absolute values. To study support recovery, we set all estimates whose absolute values are below 10^{-10} to zero, both in our estimators and the nested lasso.

In Model 1, we observe that all methods considered attain perfect classification accuracy for some value of their tuning parameter. While the non-adaptive approach is guaranteed to do so in this scenario, it is reassuring to see that the

more flexible methods can still perfectly recover this sparsity pattern.

In Model 2, we observe that our two methods outperform the nested lasso, which itself, as expected, outperforms the non-adaptive banding method. As the model becomes more challenging (from Model 2 to Model 4), the performances of all four methods start deteriorating. Interestingly, the nested lasso no longer retains its advantage over non-adaptive banding in Models 3 and 4, while the performance advantage of our methods become even more substantial.

The fact that the unweighted version of our method outperforms the weighted version stems from the fact that all models are comparatively sparse for $p = 200$, and so the heavier penalty on each row delivered by the unweighted approach recovers the support more easily than the weighted version.

2.5.2 Estimation accuracy

We proceed by comparing the estimators in terms of how far $\hat{L}$ is from L. To this end, we generate $n = 100$ random samples from the four models with $p = 50$, $p = 100$, and $p = 200$. Each method is computed with its tuning parameter selected to maximize the Gaussian likelihood on the validation data in a 5-fold cross-validation. For comparison, we report the estimation accuracy of each estimate in terms of the scaled Frobenius norm $\frac{1}{p}\left\|\hat{L} - L\right\|_F^2$, the matrix infinity norm $\left|\!\left|\!\left|\hat{L} - L\right|\!\right|\!\right|_\infty$, the spectral norm $\left|\!\left|\!\left|\hat{L} - L\right|\!\right|\!\right|_2$, and the (scaled) Kullback-Leibler loss $\frac{1}{p}\left[\mathrm{tr}(\Omega^{-1}\hat{\Omega}) - \log\det(\Omega^{-1}\hat{\Omega}) - p\right]$ (Levina et al. 2008).

The simulation is repeated 50 times, and the results are summarized in Fig-

ure 2.5 through Figure 2.8. Each figure corresponds to a model, and consists of a 4-by-3 panel layout. Each row corresponds to an error measure, and each column corresponds to a value of p.

As expected, the non-adaptive banding estimator does better than the other estimators in Model 1. In Models 2, 3, and 4, where bandwidths vary with row, our estimators and the nested lasso outperform non-adaptive banding.

A similar pattern is observed as in support recovery. As the model becomes more complex and p gets larger, the performance of the nested lasso degrades and gradually becomes worse than non-adaptive banding. By contrast, as the estimation problem becomes more difficult, the advantage in performance of our methods becomes more obvious.

We again observe that the unweighted estimator performs better than the weighted one. As shown in Section 2.4, the overall performance of our method hinges on the underlying model complexity (measured in terms of $\max_r K_r$) as well as the relative size of n and p. When n is relatively small, usually a more constrained method (like the unweighted estimator) is preferred over a more flexible method (like the weighted estimator). So in our simulation setting, it is reasonable to observe that the unweighted method works better. Note that as the underlying L becomes denser (from Model 1 to Model 4), the performance difference between the weighted and the unweighted estimator diminishes. This corroborates our discussion in the end of Section 2.4 that the performance of the unweighted estimator becomes worse when the underlying model is dense.

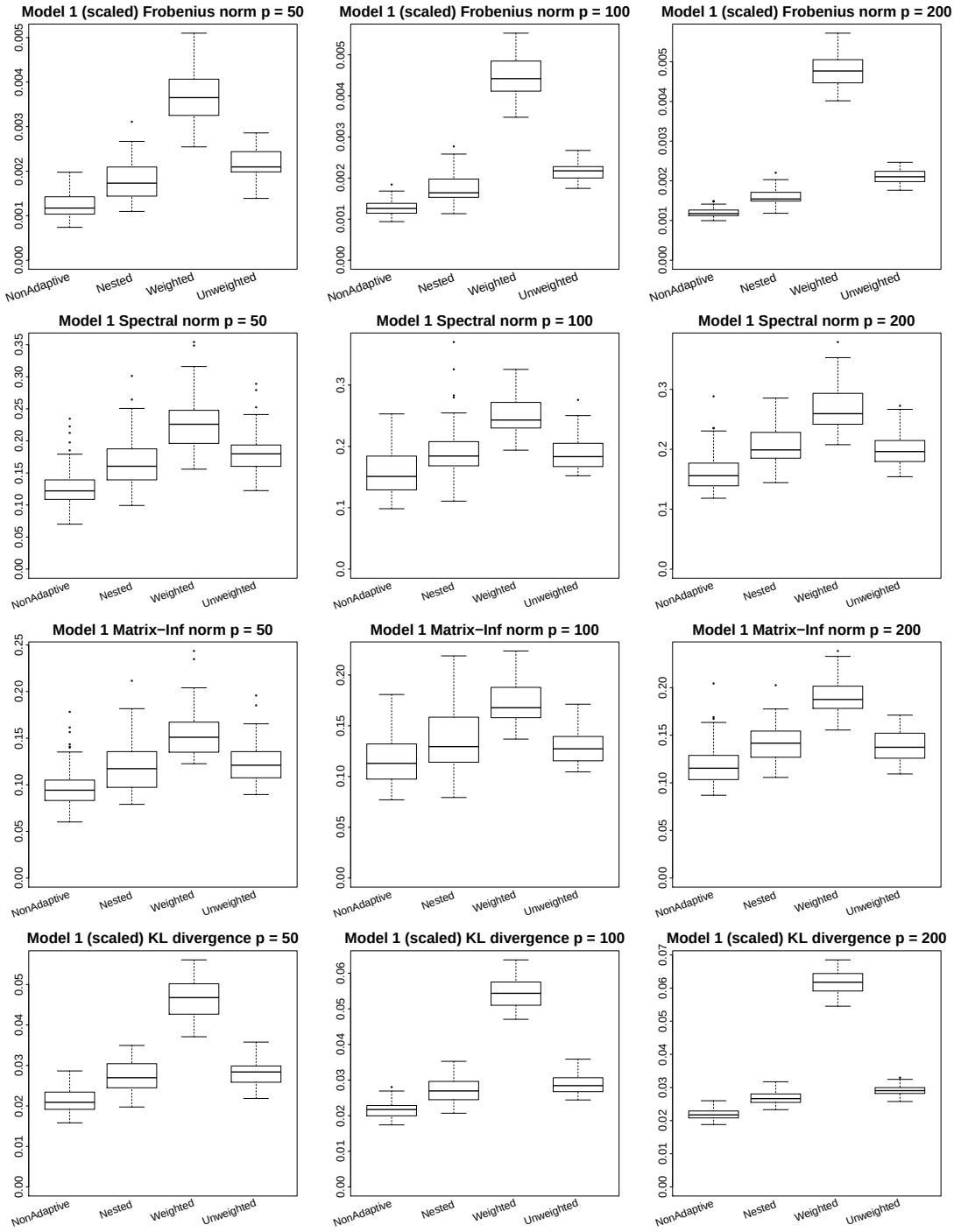

Figure 2.5: Estimation accuracy when data are generated from Model 1, which is strictly banded.

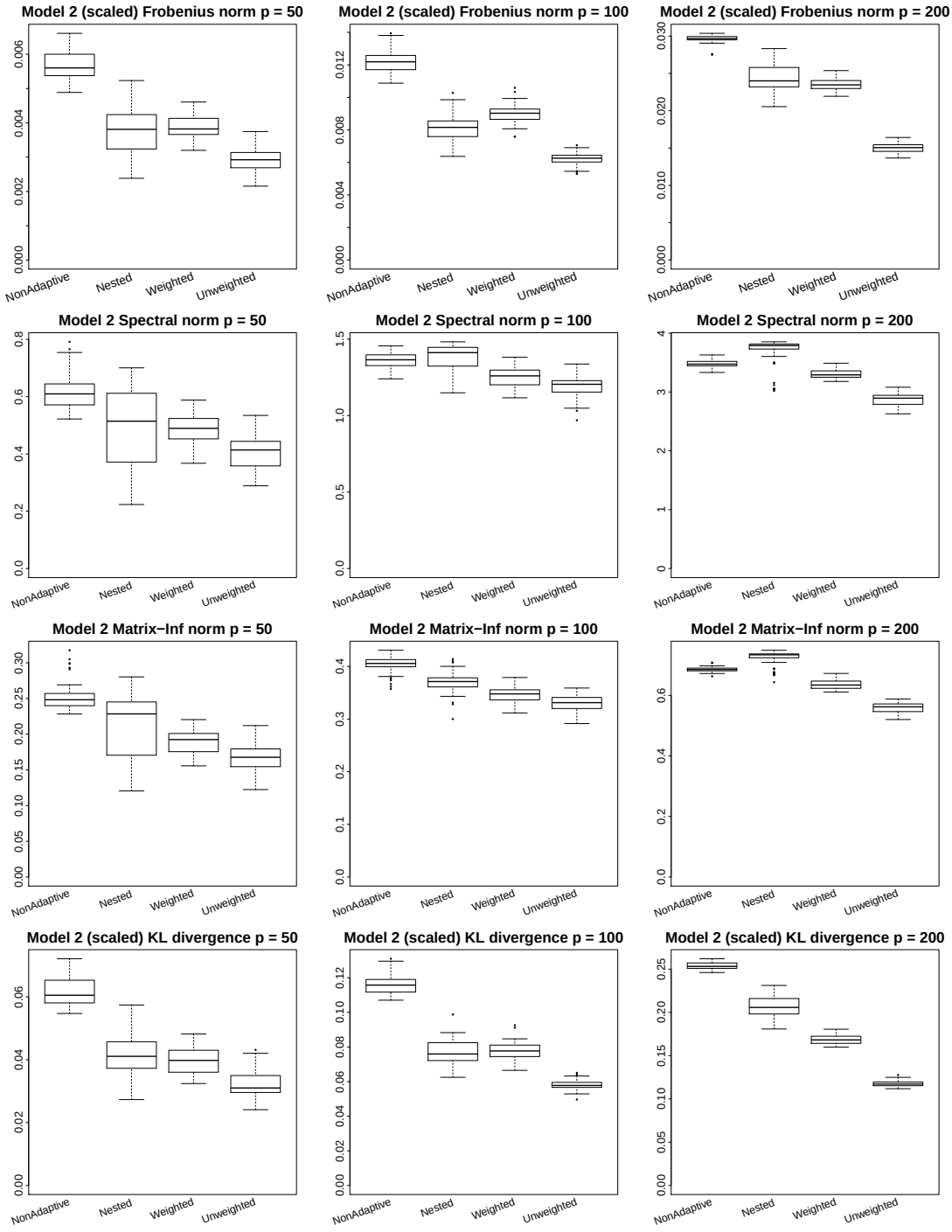

Figure 2.6: Estimation accuracy when data are generated from Model 2, which has small variable bandwidth.

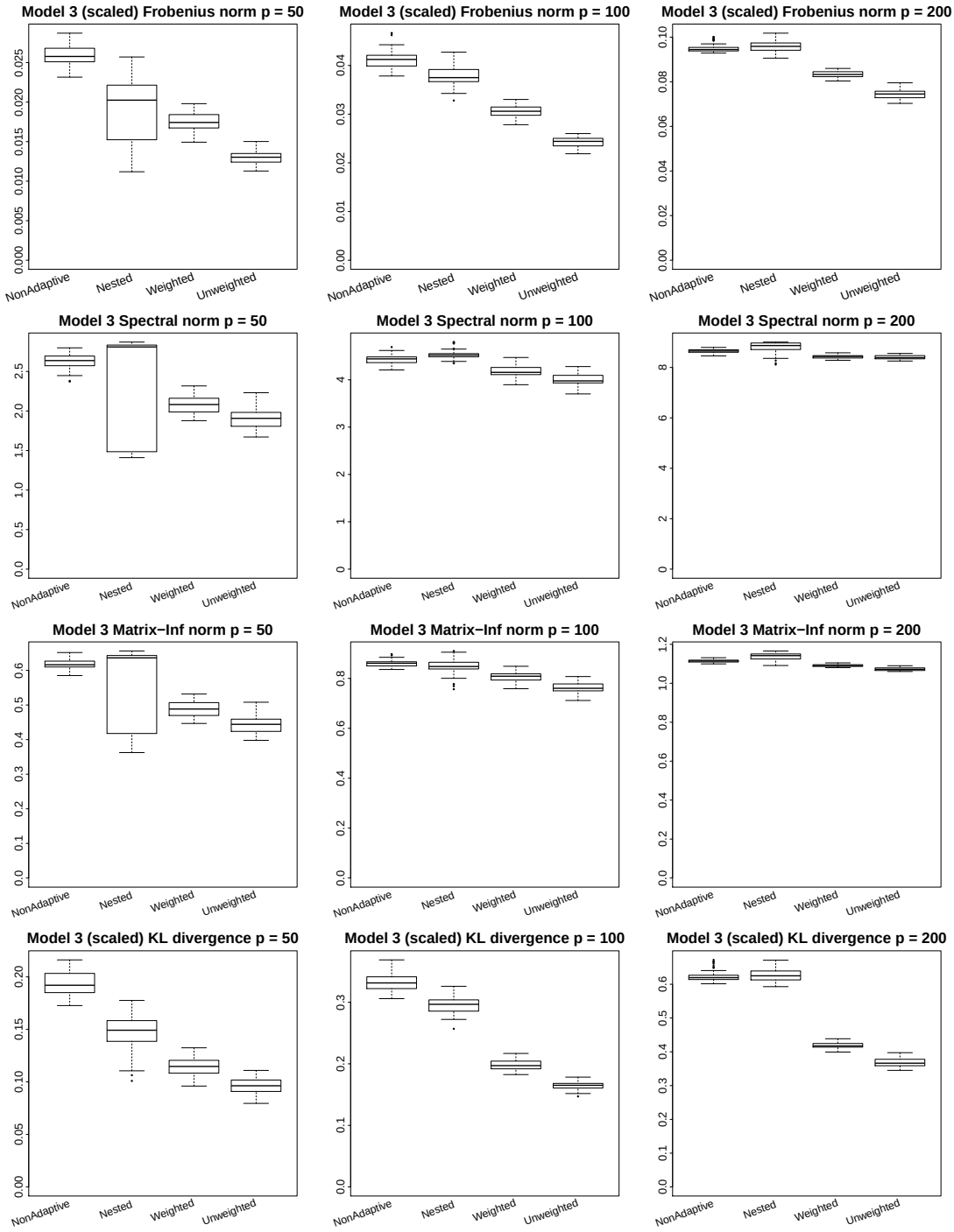

Figure 2.7: Estimation accuracy when data are generated from Model 3, which has large variable bandwidth.

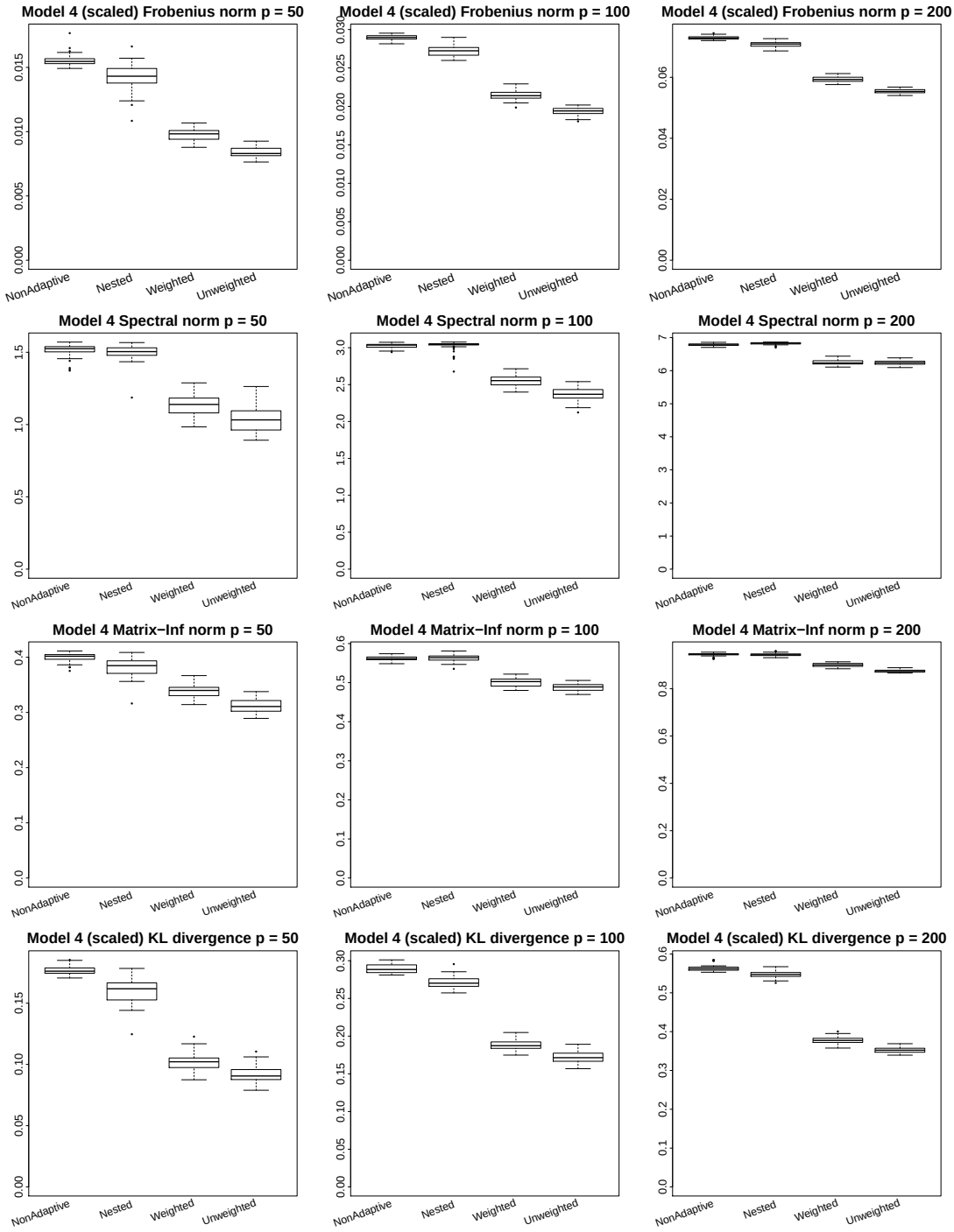

Figure 2.8: Estimation accuracy when data are generated from Model 4, which is block-diagonal.

2.6 Applications to data examples

In this section, we illustrate the practical merits of our proposed method by applying it to two data examples. We start with an application to genomic data where our method can help model the local correlations along the genome. In Section 2.6.2 we compare our method with other estimators within the context of a sound recording classification problem.

2.6.1 An application to genomic data

We consider an application of our estimator to modeling correlation along the genome. Genetic mutations that occur close together on a chromosome are more likely to be co-inherited than mutations that are located far apart (or on separate chromosomes). This leads to local correlations between genetic variants in a population. Biologists refer to this local dependence as *linkage disequilibrium* (LD). The width of this dependence is known to vary along the genome due to the variable locations of recombination hotspots, which suggests that adaptively banded estimators may be quite suitable in these contexts.

We study HapMap phase 3 data from the International HapMap project (Consortium et al. 2010). The data consist of $n = 167$ humans from the YRI (Yoruba in Ibadan, Nigeria) population, and we focus on $p = 201$ consecutive tag SNPs on chromosome 22 (after filtering out infrequent sites with minor allele frequency $\leq 10\%$).

While tag SNP data, which take discrete values $\{0, 1, 2\}$, are non-Gaussian, we argue that our estimator is still sensible to use in this case. First, the param-

eterization $\Omega = L^T L$ does not depend on the Gaussian assumption. Moreover the estimator corresponds to minimizing a penalized Bregman divergence of the log-determinant function (Ravikumar et al. 2011). Furthermore, the least-squares term in (2.5) can be interpreted as minimizing the prediction error in the linear models (2.1) while the log terms act as log-barrier functions to impose positive diagonal entries (which ensures that the resulting $\hat{L}$ is a valid Cholesky factor).

To gauge the performance of our estimator on modeling LD, we randomly split the 167 samples into training and testing sets of sizes 84 and 83, respectively. Along a path of tuning parameters with decreasing values, estimators $\hat{L}$ are computed on the training data. To evaluate $\hat{L}$ on a vector $\tilde{x}$ from the test data set, we can compute the error in predicting $\hat{L}_{rr}\tilde{x}_r$ using $-\sum_{k=1}^{r-1} \hat{L}_{r,k}\tilde{x}_k$ via (2.1) for each r, giving the error

$$\text{err}(\tilde{x}) = \frac{1}{p-1} \sum_{r=2}^{p} \left(\hat{L}_{rr}\tilde{x}_r + \sum_{k=1}^{r-1} \hat{L}_{rj}\tilde{x}_k \right)^2 . \tag{2.25}$$

This quantity (with mean and the standard deviation over test samples) is reported in Figure 2.9 for our estimator under the two weighting schemes. Recall that the quadratically decaying weights (2.7) act essentially like the ℓ_1 penalty. For numerical comparison, we also include the result of the estimator with ℓ_1 penalty, which is the *CSCS (Convex Sparse Cholesky Selection)* method proposed in Khare et al. (2016). For both the non-adaptive banding and the nested lasso methods, we found that their implementations fail to work due to the collinearity of the columns of $\mathbf{X}$.

Figure 2.9 shows that our estimators are effective in improving modeling performance over a diagonal estimator (attained when λ is sufficiently large)

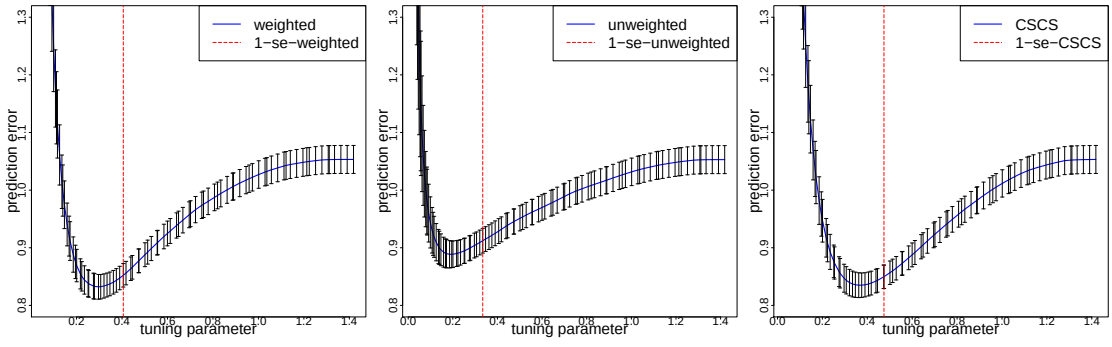

Figure 2.9: Prediction error (computed on an independent test set) of the weighted (left), unweighted (middle), and CSCS (right) estimators.

and strongly outperform the plain MLE (as evidenced by the sharp increase in prediction error as $\lambda \to 0$). As expected, the weighted estimator performs very similarly to the CSCS estimator, which uses the ℓ_1 penalty. Both of these perform better than the unweighted one. However, the sparsity pattern obtained by the two penalties are different (as shown in Figure 2.10).

In Figure 2.10 we show the recovered signed support of the weighted, unweighted, and CSCS estimators and their corresponding precision matrices. Black, gray, and white stand for positive, negative, and zero entries, respectively. Tuning parameters are chosen using the one-standard-error rule (see, e.g., Hastie et al. 2009). The r-th row of the estimated matrix $\hat{L}$ reveals the number of neighboring SNPs necessary for reliably predicting the state of the r-th SNP. Interestingly, we see some evidence of small block-like structures in $\hat{L}$, consistent with the hotspot model of recombination as previously described. This regression-based perspective to modeling LD may be a useful complement to the more standard approach, which focuses on raw marginal correlations. Finally, the sparsity recovered by the CSCS estimator, which uses the ℓ_1 penalty, is

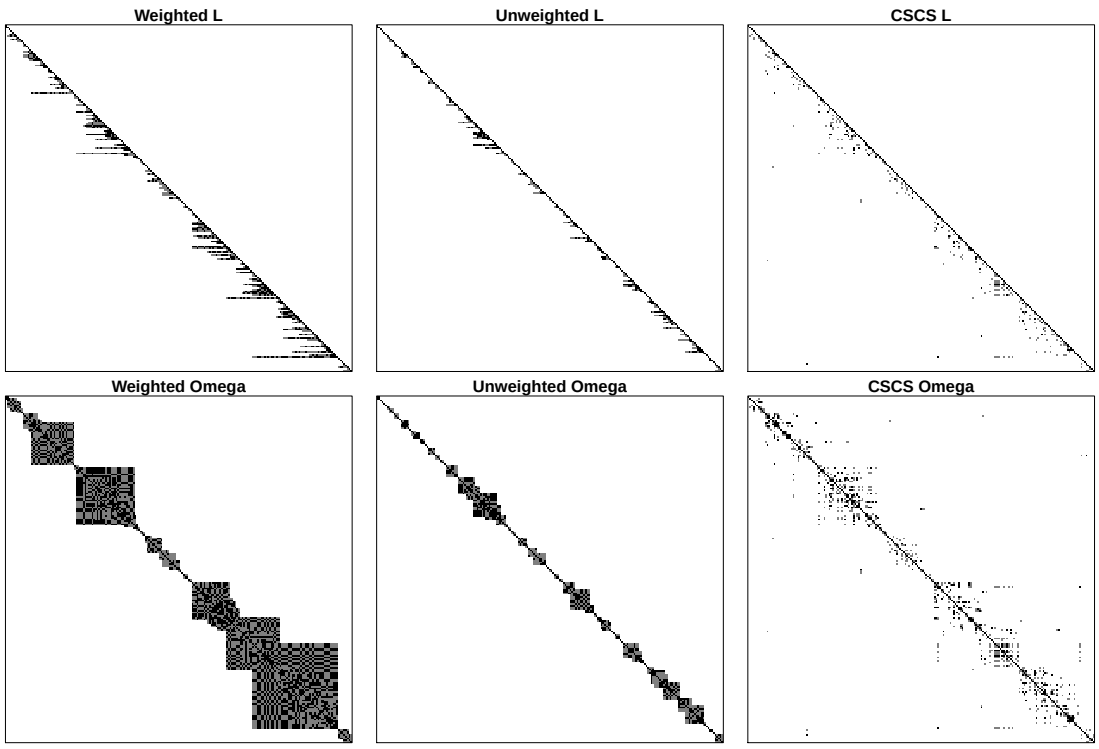

Figure 2.10: Estimates of linkage disequilibrium with tuning parameters selected by the one-standard-error rule and their corresponding precision matrix estimates.

less easily interpretable, since some entries far from the diagonal are non-zero, losing the notion of 'local'.

2.6.2 An application to phoneme classification

In this section, we develop an application of our method to a classification problem described in Hastie et al. (2009). The data contain $n = 1717$ continuous speech recordings, which are categorized into two vowel sounds: 'aa' ($n_1 = 695$) and 'ao' ($n_2 = 1022$). Each observation (x_i, y_i) has a predictor $x_i \in \mathbb{R}^p$ representing the (log) intensity of the sound across $p = 256$ frequencies and a class label

$y_i \in \{1, -1\}$. It may be reasonable to apply our method in this problem since the features are frequencies, which come with a natural ordering

In linear discriminant analysis (LDA), one models the features as multivariate Gaussian conditional on the class: $x_i|y_i = k \sim N_p(\mu^{(k)}, \Sigma)$ for $k \in \{1, -1\}$; in quadratic discriminant analysis (QDA), one allows each class to have its own covariance matrix: $x_i|y_i = k \sim N_p(\mu^{(k)}, \Sigma^{(k)})$. The LDA/QDA classification rules assign an observation $x \in \mathbb{R}^p$ to class k that maximizes $\hat{P}(y = k|x) \propto \hat{P}(x|y = k)\hat{P}(y = k)$, where the estimated probability $\hat{P}(x|y = k)$ is calculated using maximum likelihood estimates $\hat{\mu}^{(k)}$, $\hat{\Sigma}$, and $\hat{\Sigma}^{(k)}$. More precisely, in the ordered case, the resulting class k maximizes the LDA/QDA scores:

$$\begin{aligned}
\delta^{(k)}_{\text{LDA}}(x) &= x^T\hat{\Omega}\hat{\mu}^{(k)} - \frac{1}{2}(\hat{\mu}^{(k)})^T\hat{\Omega}\hat{\mu}^{(k)} + \log\hat{\pi}^{(k)} \\
&= (\hat{L}x)^T\hat{L}\hat{\mu}^{(k)} - \frac{1}{2}\left\|\hat{L}\hat{\mu}^{(k)}\right\|_2^2 + \log\hat{\pi}^{(k)} \qquad (2.26) \\
\delta^{(k)}_{\text{QDA}}(x) &= x^T\hat{\Omega}^{(k)}\hat{\mu}^{(k)} - \frac{1}{2}(\hat{\mu}^{(k)})^T\hat{\Omega}^{(k)}\hat{\mu}^{(k)} + \log\hat{\pi}^{(k)} \\
&= (\hat{L}^{(k)}x)^T\hat{L}^{(k)}\hat{\mu}^{(k)} - \frac{1}{2}\left\|\hat{L}^{(k)}\hat{\mu}^{(k)}\right\|_2^2 + \log\hat{\pi}^{(k)}. \qquad (2.27)
\end{aligned}$$

Note that it is the precision matrix, not the covariance matrix, that is used in the above scores. In the setting where $p > n$, the MLE of Ω or $\Omega^{(k)}$ does not exist. A regularized estimate of precision matrix that exploits the natural ordering information can be helpful in this setting.

To demonstrate the use of our estimator in the high-dimensional setting, we randomly split the data into two parts, with 10% of the data assigned to the training set and the remaining 90% of the data assigned to the test set. On the training set, we use 5-fold cross-validation to select the tuning parameter minimizing misclassification error on the validation data. The estimates $\hat{L}$ and $\hat{L}^{(k)}$ are then plugged into (2.26) and (2.27) along with $\hat{\mu}^{(k)} = \sum_{i\in\text{class }k} x_i/n^{(k)}$ and

	Unweighted	Weighted	Nested Lasso	Non-adaptive	CSCS
LDA	0.271	0.246	0.250	0.268	0.245
QDA	0.232	0.256	0.221	0.246	0.267

Table 2.1: Average test data classification error rate of discriminant analysis of phoneme data

$\hat{\pi}^{(k)} = n^{(k)}/n_{\text{train}}$ to calculate the misclassification error in the test set. For comparison, we also include non-adaptive banding, the nested lasso, and CSCS. We compute the classification error (summarized in Table 2.1), averaged over 10 random train-test splits.

We first observe that, in general, the adaptive methods perform better than the non-adaptive one (which assumes a fixed bandwidth). It is again found that the performance of the weighted estimator is very similar to the one using ℓ_1 penalty (i.e., the CSCS method). And our results are comparable to the nested lasso both in LDA and QDA. Interestingly, we find that the weighted estimator does better in LDA while the unweighted estimator performs better in QDA. The reason, we suspect, is that QDA requires the estimation of more parameters than LDA and therefore favors more constrained methods like the unweighted estimator, which more strongly discourages non-zeros from being far from the diagonal than the weighted one.

An `R` (R Core Team 2017) package, named `varband`, is available on `CRAN`, implementing our estimator. The estimation is very fast with core functions coded in `C++`, allowing us to solve large-scale problems in substantially less time than is possible with the `R`-based implementation of the nested lasso.

Acknowledgement

We thank Kshitij Khare for a useful discussion in which he pointed us to the parametrization in terms of L. We thank Adam Rothman for providing `R` code for the non-adaptive banding and the nested lasso methods and Amy Williams for useful discussions about linkage disequilibrium. We also thank three referees and an action editor for helpful comments on an earlier manuscript. This work was supported by NSF DMS-1405746.

CHAPTER 3

ESTIMATING THE ERROR VARIANCE IN A HIGH-DIMENSIONAL LINEAR MODEL

Portions of this chapter were available in Yu & Bien (2017a)

3.1 Introduction

The linear model

$$\mathbf{y} = \mathbf{X}\beta^* + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I_n), \tag{3.1}$$

is one of the most fundamental models in statistics. It describes the relationship between a vector $\mathbf{y} \in \mathbb{R}^n$ of n independent observations of a response variable and a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of n observations of p features. The unknown parameters of this model are the vector of coefficients $\beta^* \in \mathbb{R}^p$, which expresses how $\mathbf{y}$ relates to $\mathbf{X}$, and the error variance σ^2, which captures the noise level or extent to which $\mathbf{y}$ cannot be predicted from $\mathbf{X}$: The vector $\varepsilon \in \mathbb{R}^n$ consists of independently and identically distributed zero-mean Gaussian errors with variance σ^2. When $p \gg n$, estimating β^* is a challenging, well-studied problem. Perhaps the most common method in this setting is the *lasso* (Tibshirani 1996), which assumes that β^* is sparse and solves the following convex optimization problem:

$$\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(n^{-1} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + 2\lambda \|\beta\|_1 \right). \tag{3.2}$$

Over the past decade, an extensive literature has emerged studying the properties of $\hat{\beta}_\lambda$ from both computational (see, e.g., Hastie et al. 2015) and theoretical (see, e.g., Bühlmann & Van De Geer 2011) perspectives.

Compared to the vast amount of work on estimating β^*, relatively little attention has been paid to the problem of estimating σ^2. Nonetheless, reliable estimation of σ^2 is important for quantifying the uncertainty in estimating β^*. A series of recent advances in high-dimensional inference (Bühlmann 2013, Zhang & Zhang 2014, Van de Geer et al. 2014, Lockhart et al. 2014, Javanmard & Montanari 2014, Lee et al. 2016, Tibshirani et al. 2016, Taylor & Tibshirani 2017, Ning & Liu 2017, etc.) may very well be the determining factor for the widespread adoption of the lasso and the related methods in fields where p-values and confidence intervals are required. Point estimates without accompanying inferential statements are distrusted and disregarded in these areas. Estimating σ^2 reliably in finite sample is crucial.

If β^* were known, then the optimal estimator for σ^2 would of course be $n^{-1}\|\mathbf{y} - \mathbf{X}\beta^*\|_2^2 = n^{-1}\|\varepsilon\|_2^2$. Thus, a naive estimator for σ^2 based on an estimator $\hat{\beta}$ of β^* would be

$$\hat{\sigma}^2_{\text{naive}} = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2. \tag{3.3}$$

However, a simple calculation in the classical $n > p$ setting shows that such an estimator is biased downward: a least-squares oracle with knowledge of the true support $S = \{j : \beta^*_j \neq 0\}$ scales this to give an unbiased estimator:

$$\hat{\sigma}^2_{\text{oracle}} = \frac{1}{n - |S|}\|\mathbf{y} - \mathbf{X}_S\mathbf{X}_S^+\mathbf{y}\|_2^2, \tag{3.4}$$

where $\mathbf{X}_S$ is a sub-matrix of $\mathbf{X}$ with columns indexed by S and $\mathbf{X}_S^+$ is its pseudoinverse. Many papers in this area discuss the difficulty of estimating σ^2 and warn of the perils of underestimating it: if σ^2 is underestimated then one gets anti-conservative confidence intervals, which are highly undesirable (Tibshirani et al. 2018).

Reid et al. (2016) carry out an extensive review and simulation study of several estimators of σ^2 (Fan et al. 2012, Sun & Zhang 2012, Dicker 2014), and they devote special attention to studying the estimator

$$\hat{\sigma}_R^2 = \frac{1}{n - \hat{s}_\lambda} \|\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\|_2^2, \tag{3.5}$$

where $\hat{\beta}_\lambda$ is as in (3.2), with λ selected using a cross-validation procedure, and $\hat{s}_\lambda$ is the number of nonzero elements in $\hat{\beta}_\lambda$. They show that (3.5) has promising performance in a wide range of simulation settings and provide an asymptotic theoretical understanding of the estimator in the special case where $\mathbf{X}$ is an orthogonal matrix.

While intuition from (3.4) suggests that (3.5) is a quite reasonable estimator when S can be well recovered, it also points to the question of how well the estimator will perform when S is not well recovered by the lasso. The conditions required for the lasso to recover S are much stricter than the conditions needed for it to do well in prediction (see, e.g., Van de Geer & Bühlmann 2009). The scale factor $(n - \hat{s}_\lambda)^{-1}$ used in $\hat{\sigma}_R^2$ means that this approach depends not just on the predicted values of the lasso, $\mathbf{X}\hat{\beta}_\lambda$, but on the magnitude of the set of nonzero elements in $\hat{\beta}_\lambda$. Indeed, we find that in situations where recovering S is known to be challenging, $\hat{\sigma}_R^2$ tends to yield less favorable empirical performance. The theoretical development in Reid et al. (2016) sidesteps this complication by working in an asymptotic regime in which $\hat{\sigma}_R^2$ behaves like the naive estimator (3.3). To understand the finite-sample performance of $\hat{\sigma}_R^2$ would require considering the behavior of the random variable $\hat{s}_\lambda$. Clearly, when $\hat{s}_\lambda \approx n$, even small fluctuations in $\hat{s}_\lambda$ can lead to large fluctuations in $\hat{\sigma}_R^2$. Finally, from a practical standpoint, computing $\hat{s}_\lambda$ is a numerically sensitive operation in that it requires the choice of a threshold size for calling a value numerically zero (and the assurance that one has solved the problem to sufficient precision).

Based on these observations, we propose in this paper a completely different approach to estimating σ^2. The basic premise of our framework is that when both β^* and σ^2 are unknown, it is convenient to formulate the penalized log-likelihood problem in terms of

$$\phi = \sigma^{-2}, \qquad \theta = \sigma^{-2}\beta, \tag{3.6}$$

the natural parameters of the Gaussian multiparameter exponential family with unknown mean and variance. Whereas the negative Gaussian log-likelihood is not jointly convex in the (β, σ) parameterization (in fact, it is nonconvex in σ), in the natural parameterization the negative log-likelihood is jointly convex in (ϕ, θ).

We penalize this negative log-likelihood with an ℓ_1-norm on the natural parameter θ and call this new estimator the *natural lasso*. We show in Section 3.3 that the resulting error variance estimator can in fact be very simply expressed as the minimizing value of the regular lasso problem (3.2):

$$\hat{\sigma}^2_\lambda = \min_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + 2\lambda \|\beta\|_1 \right). \tag{3.7}$$

Observing that the first term is $\hat{\sigma}^2_{\text{naive}}$, we directly see that the natural lasso counters the naive method's downward bias through an *additive* correction; this is in contrast to $\hat{\sigma}^2_R$'s reliance on a (sometimes unstable) multiplicative correction. Computing (3.7) is clearly no harder than solving a lasso and, unlike $\hat{\sigma}^2_R$, does not require determining a threshold for deciding which coefficient estimates are numerically zero. Furthermore, we establish finite sample bounds on the mean squared error that hold without making any assumptions on the design matrix $\mathbf{X}$. Our theoretical analysis suggests a second approach that is also based on the natural parameterization. The theory that we develop for this method, which

we call the *organic lasso*, relies on weaker assumptions. We find that both methods have competitive empirical performance relative to $\hat{\sigma}_R^2$ and show particular strength in settings in which support recovery is known to be challenging.

In addition, when $\hat{\beta}$ in (3.3) is taken to be the standard lasso or the square-root/scaled lasso estimator (Belloni et al. 2011, Sun & Zhang 2012), we cannot find in previous literature an indication of whether $\hat{\sigma}^2_{\text{naive}}$ can match the same rate of convergence as our estimators when one does not place assumptions on the design matrix X. In this paper, we show that this is in fact the case, thus providing a fuller story about the problem of estimating the error variance in high-dimensional linear models

3.2 Natural parameterization

The negative log-likelihood function in (3.1) is (up to a constant)

$$\mathcal{L}\left(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}\right) = \frac{n}{2} \log \sigma^2 + \frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2}.$$

When σ^2 is known, the σ dependence can be ignored, leading to the standard least-squares criterion; however, when σ is unknown, performing a full minimization of the penalized negative log-likelihood amounts to solving a nonconvex optimization problem even with a convex penalty.

The nonconvexity of the Gaussian negative log-likelihood in its variance (or, more generally, covariance matrix) is a well-known difficulty (Bien & Tibshirani 2011). In this context, working instead with the *inverse* covariance matrix is common (Yuan & Lin 2007, Banerjee et al. 2008, Friedman et al. 2008). We take an analogous approach here, considering the natural parameterization (3.6) of

the Gaussian multiparameter exponential family with unknown variance,

$$\mathcal{L}\left(\phi^{-1}\theta, \phi^{-1} | \mathbf{X}, \mathbf{y}\right) = -\frac{n}{2}\log\phi + \frac{1}{2}\phi\left\|\mathbf{y} - \mathbf{X}\frac{\theta}{\phi}\right\|_2^2 = -\frac{n}{2}\log\phi + \phi\frac{\|y\|_2^2}{2} - \mathbf{y}^T\mathbf{X}\theta + \frac{\|\mathbf{X}\theta\|_2^2}{2\phi}.$$

Observing that attaining sparsity in θ is equivalent to attaining sparsity in β, we propose the following penalized maximum log-likelihood estimator:

$$\left(\hat{\theta}_\lambda, \hat{\phi}_\lambda\right) \in \underset{\phi>0,\ \theta}{\arg\min}\left(-\frac{1}{2}\log\phi + \phi\frac{\|y\|_2^2}{2n} - \frac{1}{n}\mathbf{y}^T\mathbf{X}\theta + \frac{\|\mathbf{X}\theta\|_2^2}{2n\phi} + \lambda\Omega(\theta, \phi)\right) \tag{3.8}$$

for a convex penalty $\Omega(\theta, \phi)$ that induces sparsity in θ. We will focus on $\Omega(\theta, \phi) = \|\theta\|_1$ in Section 3.3 and $\Omega(\theta, \phi) = \|\theta\|_1^2/\phi$ in Section 3.4. This problem is jointly convex in (θ, ϕ). While this is a general property of exponential families (due to the convexity of the cumulant generating function), we can see it in this special case because of the convexity of $-\log$ and the convexity of the "quadratic-over-linear" function (Boyd & Vandenberghe 2004). Given a solution to (3.8), we can reverse (3.6) to get estimators for σ^2 and β^*:

$$\tilde{\sigma}_\lambda^2 = \hat{\phi}_\lambda^{-1}, \qquad \tilde{\beta}_\lambda = \hat{\phi}_\lambda^{-1}\hat{\theta}_\lambda. \tag{3.9}$$

Before proceeding with an analysis of the estimator (3.9) with specific choices of $\Omega(\theta, \phi)$, we point out a similarity between our method and that of Städler et al. (2010), who consider a different convexifying reparameterization of the Gaussian log-likelihood, using $\rho = \sigma^{-1}$ and $\gamma = \sigma^{-1}\beta$. They put an ℓ_1-norm penalty on γ (which has the same sparsity pattern as β) and solve

$$\min_{\rho>0,\gamma}\left(-\log\rho + \frac{1}{2n}\|\rho\mathbf{y} - \mathbf{X}\gamma\|_2^2 + \lambda\|\gamma\|_1\right). \tag{3.10}$$

Sun & Zhang (2010) give an asymptotic analysis of the solution to (3.10) under a compatibility condition. A modification of this problem (Antoniadis 2010) gives the scaled lasso (Sun & Zhang 2012), which is known to be equivalent to

the square-root lasso (Belloni et al. 2011):

$$\tilde{\beta}_{\text{SQRT}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{\sqrt{n}} \|\mathbf{y} - \mathbf{X}\beta\|_2 + \lambda \|\beta\|_1 \right), \qquad \tilde{\sigma}^2_{\text{SQRT}} = \frac{1}{n} \left\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right\|_2^2. \tag{3.11}$$

With the same parameterization (ρ, γ), Dalalyan & Chen (2012) propose the scaled Dantzig selector under the assumption of fused sparsity. Under the restricted eigenvalue condition, they establish the same rate of convergence in estimating the error variance as the fast prediction error rate of the standard lasso.

3.3 The natural lasso estimator of error variance

We first propose the natural lasso, which is the solution to (3.8) with $\Omega(\theta, \phi) = \|\theta\|_1$. One might think that solving the natural lasso would involve a specialized algorithm. The following proposition shows, remarkably, that this is not the case.

Proposition 7. *The natural lasso estimator $(\tilde{\beta}_\lambda, \tilde{\sigma}^2_\lambda)$ defined in* (3.9)*, where $(\hat{\theta}_\lambda, \hat{\phi}_\lambda)$ is a solution to* (3.8) *with $\Omega(\theta, \phi) = \|\theta\|_1$, satisfies the following properties:*

1. *$\tilde{\beta}_\lambda = \hat{\beta}_\lambda$, a solution to the standard lasso* (3.2)*;*
2. *$\tilde{\sigma}^2_\lambda = \hat{\sigma}^2_\lambda$, the standard lasso's optimal value* (3.7)*.*

Furthermore, $\hat{\sigma}^2_\lambda = (\|\mathbf{y}\|_2^2 - \|\mathbf{X}\hat{\beta}_\lambda\|_2^2)/n$.

The proof of this proposition and all theoretical results that follow can be found in Appendix B. Thus, to get the natural lasso estimator of (β^*, σ^2), one

simply solves the standard lasso (3.2) and returns a solution and the minimal value.

An attractive property of the natural lasso estimator $\hat{\sigma}^2_\lambda$ is the relative ease with which one can prove bounds about its performance. Since $\hat{\sigma}^2_\lambda$ is the optimal value of the lasso problem, the objective value at any vector β provides an upper bound on $\hat{\sigma}^2_\lambda$. Likewise, any dual feasible vector provides a lower bound on $\hat{\sigma}^2_\lambda$. These considerations are used to prove the following lemma, which shows that for a suitably chosen λ, the natural lasso variance estimator gets close to the oracle estimator of σ^2.

Lemma 8. *If $\lambda \geq n^{-1}\|\mathbf{X}^T\varepsilon\|_\infty$, then $|\hat{\sigma}^2_\lambda - n^{-1}\|\varepsilon\|_2^2| \leq 2\lambda\|\beta^*\|_1$.*

The result above is "deterministic" in that it does not rely on any statistical assumptions or arguments. The next result adds such considerations to give a mean squared error bound for the natural lasso.

Theorem 9. *Suppose that each column $\mathbf{X}_j$ of the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has been scaled so that $\|\mathbf{X}_j\|_2^2 = n$ for all $j = 1, \ldots, p$, and assume that $\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$. Then, for any constant $M > 1$, the natural lasso estimator* (3.7) *with $\lambda = \sigma(2Mn^{-1}\log p)^{1/2}$ satisfies the following relative mean squared error bound:*

$$\mathrm{E}\left\{\left(\frac{\hat{\sigma}^2_\lambda}{\sigma^2} - 1\right)^2\right\} \leq \left\{\left(8M + 8\frac{p^{1-8M}}{\log p}\right)^{1/2} \frac{\|\beta^*\|_1}{\sigma}\left(\frac{\log p}{n}\right)^{1/2} + \left(\frac{2}{n}\right)^{1/2}\right\}^2.$$

Corollary 10.

$$\mathrm{E}\left|\frac{\hat{\sigma}^2_\lambda}{\sigma^2} - 1\right| = O\left\{\frac{\|\beta^*\|_1}{\sigma}\left(\frac{\log p}{n}\right)^{1/2}\right\}. \tag{3.12}$$

Proof. This follows from Jensen's inequality. □

Remark 11. *Theorem 9 can be easily generalized to the case where the i.i.d. zero-mean error ε_i with variance σ^2 is sub-Gaussian or sub-exponential. A high probability bound*

can be obtained for ε_i with bounded polynomial moments. In particular, for any $m \geq 3$, if $\mathrm{E}(|\varepsilon_i|^m) \leq (m!)^{-1} 2K^{m-2}$ for some $K > 0$, and if each column X_j is scaled so that $\sum_{i=1}^n X_{ij}^m = n$ for $j = 1, \ldots, p$, then with $\lambda = 4K\sigma n^{-1/2}(\log p)^{1/2}$ we have that

$$\left| \hat{\sigma}_\lambda^2 - \frac{\|\varepsilon\|_2^2}{n} \right| = O\left\{ \sigma \|\beta^*\|_1 \left(\frac{\log p}{n} \right)^{1/2} \right\}$$

holds with probability greater than $1 - p^{-1}$.

To put Theorem 9 in context, we devote the remainder of this section to considering what bounds are available for other methods for estimating σ^2. Bayati et al. (2013) propose an estimator of σ^2 based on estimating the mean squared error of the lasso. They show that their estimator of σ^2 is asymptotically consistent with fixed p as $n \to \infty$. In contrast, we provide finite sample results and these include the $p \gg n$ case. Also, the consistency result in Bayati et al. (2013) is based on the assumption of independent Gaussian features (and in extending this to the case of correlated Gaussian features, the authors invoke a conjecture). In comparison, (3.12) is essentially free of assumptions on the design matrix.

The natural lasso also compares favorably to the method-of-moments-based estimator of Dicker (2014) in terms of mean squared error bounds. In particular, Dicker (2014) establishes a $O_P[(\tau^2/\sigma^2 + 1)\{(p+n)/n^2\}^{1/2}]$ relative mean squared error rate, where $\tau^2 = \|\Sigma^{-1/2}\beta^*\|_2^2$ and Σ is the covariance of features X. This rate can be much slower for large p.

Notably, the mean squared error bound in Theorem 9 does not put any assumption on $\mathbf{X}$, β^*, or σ^2. In this sense, the result is analogous to a "*slow rate*" bound (Rigollet & Tsybakov 2011, Dalalyan et al. 2017), which appears in the lasso prediction consistency context. While it is well known (Sun & Zhang 2012) or can be easily verified that under stronger conditions (i.e., compatibility or re-

stricted eigenvalue conditions) the naive estimator (3.3) based on the lasso and $\tilde{\sigma}^2_{\text{SQRT}}$ in (3.11) attain a faster rate, $O(|S|n^{-1}\log p)$, it is natural to ask whether these two estimators also attain a rate bound as in (3.12) when the conditions on $\mathbf{X}$ are not assumed. The following two results give an affirmative answer to this question.

Proposition 12. *Under the conditions of Theorem 9, the naive estimator* (3.3) *based on the lasso estimator* $\hat{\beta}_\lambda$ *with* $\lambda = 4\sigma(n^{-1}\log p)^{1/2}$ *has the following bound with probability greater than* $1 - p^{-1}$*:*

$$\left|\hat{\sigma}^2_{\text{naive}} - n^{-1}\|\varepsilon\|_2^2\right| \leq 16\sigma\|\beta^*\|_1\left(\frac{\log p}{n}\right)^{1/2}. \tag{3.13}$$

Relatedly, Chatterjee & Jafarov (2015) also consider a setting with no assumptions on X and derive an error bound $O\{\|\beta^*\|_1^{1/2}(n^{-1}\log p)^{1/4}\}$ for (3.3) for a lasso estimator $\hat{\beta}_\lambda$ with λ in (3.2) selected using a cross-validation procedure.

Lederer et al. (2016) derive a slow rate bound for the prediction error of the square root lasso. They show (in Lemma 2.1) that there exists a value of λ for which $\lambda = 3n^{-1/2}\|\mathbf{X}^T\varepsilon\|_\infty\|\mathbf{y}-\mathbf{X}\tilde{\beta}_{\text{SQRT}}\|_2^{-1}$ and bound $\|X\tilde{\beta}_{\text{SQRT}}-X\beta^*\|_2^2$ at this value. The following result establishes the high-dimensional consistency of $\tilde{\sigma}^2_{\text{SQRT}}$ under no assumptions on X.

Proposition 13. *Under the conditions of Theorem 9, for the square-root/scaled lasso estimator* $\tilde{\sigma}^2_{\text{SQRT}}$ *in* (3.11) *based on* $\tilde{\beta}_{\text{SQRT}}$ *with* $\lambda = 3n^{-1/2}\|\mathbf{X}^T\varepsilon\|_\infty\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\|_2^{-1}$ *has the following bound with probability greater than* $1 - p^{-1}$*:*

$$\left|\tilde{\sigma}^2_{\text{SQRT}} - \frac{1}{n}\|\varepsilon\|_2^2\right| \leq 12\sigma\|\beta^*\|_1\left(\frac{\log p}{n}\right)^{1/2}. \tag{3.14}$$

We see the rate of the natural lasso in (3.12) matches (up to a constant factor) the rates (3.13) and (3.14). The values of λ used in Propositions 12 and 13

are larger than would be necessary for standard prediction error bounds; we learned of this technique from Irina Gaynanova (Gaynanova n.d.), and it is key to the proofs of the two propositions.

3.4 The organic lasso estimator of error variance

3.4.1 Method formulation

In practice, the value of the regularization parameter λ in (3.7) may be chosen via cross-validation; however, Theorem 9 has a regrettable theoretical shortcoming: it requires using a value of λ that itself depends on σ, the very quantity that we are trying to estimate! This is a well-known theoretical limitation of the lasso and related methods that motivated the square-root/scaled lasso. In this section, we propose a second new method, which retains the natural parameterization, but remedies the natural lasso's theoretical shortcoming by using a modified penalty. We define the organic lasso as a solution to (3.8) with $\Omega(\theta, \phi) = \|\theta\|_1^2/\phi$, i.e.,

$$\left(\check{\theta}_\lambda, \check{\phi}_\lambda\right) = \underset{\phi>0,\ \theta}{\arg\min} \left(-\frac{1}{2}\log\phi + \phi\frac{\|y\|_2^2}{2n} - \frac{1}{n}\mathbf{y}^T\mathbf{X}\theta + \frac{\|\mathbf{X}\theta\|_2^2}{2n\phi} + \lambda\frac{\|\theta\|_1^2}{\phi} \right). \tag{3.15}$$

We observe that the penalty $\|\theta\|_1^2/\phi$ is jointly convex in (ϕ, θ) since it can be expressed as $g(h(\theta), \phi)$ where $h(\theta) = \|\theta\|_1$ is convex and $g(x, \phi) = x^2/\phi$ is a jointly convex function that is strictly increasing in x for $x \geq 0$ (Boyd & Vandenberghe 2004).

Given a solution to the above problem, we can reverse (3.6) to give the or-

ganic lasso estimators of (β^*, σ^2):

$$\check{\beta}_\lambda = \check{\phi}_\lambda^{-1}\check{\theta}_\lambda, \qquad \check{\sigma}_\lambda^2 = \check{\phi}_\lambda^{-1}.$$

In direct analogy to the natural lasso, the following proposition shows that we can find $\check{\sigma}_\lambda^2$ and $\check{\beta}_\lambda$ without actually solving (3.15).

Proposition 14. *The organic lasso estimators $(\check{\beta}_\lambda, \check{\sigma}_\lambda^2)$ correspond to the solution and minimal value of an ℓ_1^2-penalized least-squares problem:*

$$\check{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(\frac{1}{n}\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + 2\lambda\|\beta\|_1^2\right); \tag{3.16}$$

$$\check{\sigma}_\lambda^2 = \min_{\beta\in\mathbb{R}^p}\left(\frac{1}{n}\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + 2\lambda\|\beta\|_1^2\right). \tag{3.17}$$

Thus, to compute the organic lasso estimator, one simply solves a penalized least squares problem, where the penalty is the square of the ℓ_1 norm. This can be thought of as the exclusive lasso with a single group (Zhou et al. 2010, Campbell et al. 2017). We show in the next section that solving this problem is no harder than solving a standard lasso problem.

One readily sees the connection of the organic lasso to the square-root lasso (3.11): to get (3.17), one takes squares of both the loss and the ℓ_1 penalty of (3.11). However, their origins are actually different in nature: the organic lasso is a maximum of the Gaussian log-likelihood with a scale-equivariant sparsity inducing penalty under parameterization (3.6), while (3.11) minimizes the ℓ_1-penalized Huber concomitant loss function (Antoniadis 2010, Sun & Zhang 2012).

3.4.2 Algorithm

Coordinate descent is easy to implement and has steadily maintained its place as a start-of-the-art approach for solving lasso-related problems (Friedman et al. 2007). For coordinate descent to work, one typically verifies separability in the non-smooth part of the objective function (Tseng 2001). However, the ℓ_1^2 penalty in (3.16) is not separable in the coordinates of β. Lorbert et al. (2010) propose a coordinate descent algorithm to solve the Pairwise Elastic Net (PEN) problem, a generalization of (3.16), and a proof of the convergence of the algorithm is given in Lorbert (2012). In Algorithm 3, we give a coordinate descent algorithm specific to solving (3.16). The `R` package `natural` (Yu 2017) provides a `C` implementation of Algorithm 3.

Algorithm 3: A coordinate descent algorithm to solve (3.16)

Require: Initial estimate $\beta^{(0)} \in \mathbb{R}^p$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, and $\lambda > 0$.

Set $\beta \leftarrow \beta^{(0)}$ and $\mathbf{r} \leftarrow \mathbf{y} - \mathbf{X}\beta$

for $j = 1, \ldots, p; 1, \ldots, p; \ldots$ (until convergence) **do**:

$\beta_j^{\text{new}} \leftarrow (2\lambda + \left\|\mathbf{X}_j\right\|_2^2 / n)^{-1} \mathcal{S}(\mathbf{X}_j^T \mathbf{r}/n + \left\|\mathbf{X}_j\right\|_2^2 \beta_j / n, 2\lambda \|\beta_{-j}\|_1)$

$\mathbf{r} \leftarrow \mathbf{r} + \mathbf{X}_j \beta_j - \mathbf{X}_j \beta_j^{\text{new}}$

$\beta_j \leftarrow \beta_j^{\text{new}}$

return β.

Each coordinate update is $O(n)$, where $\mathcal{S}(a, b) = \text{sgn}(a)(|a| - b)_+$ is the soft-threshold operator. Empirically Algorithm 3 is found to be essentially as fast as solving a lasso problem. Theorem C.3.9 in Lorbert (2012) shows that, for any initial estimate $\beta^{(0)} \in \mathbb{R}^p$, every limit point of Algorithm 3 is an optimal point of the objective function of (3.16). This implies that the ℓ_1^2 penalty, although

not separable, is well enough behaved that any point that is minimum in every coordinate of the objective function in (3.16) is indeed a global minimum.

3.4.3 Theoretical results

A first indication that the organic lasso may succeed where the natural lasso falls short is in terms of scale equivariance. As the design $\mathbf{X}$ is usually standardized to be unitless, scale equivariance in this context refers to the effect of scaling $\mathbf{y}$.

Proposition 15. *The organic lasso is scale equivariant, i.e., for any $t > 0$,*

$$\check{\beta}_\lambda(t\mathbf{y}) = t\check{\beta}_\lambda(\mathbf{y}), \qquad \check{\sigma}_\lambda(t\mathbf{y}) = t\check{\sigma}_\lambda(\mathbf{y}).$$

Scale equivariance is a property associated with the ability to prove results in which the tuning parameter λ does not depend on σ. For example, the square-root/scaled lasso (3.11) is scale equivariant while the lasso (and thus the natural lasso) is not. In particular, $\hat{\beta}_\lambda(t\mathbf{y}) \neq t\hat{\beta}_\lambda(\mathbf{y})$, and $\hat{\sigma}_\lambda(t\mathbf{y}) \neq t\hat{\sigma}_\lambda(\mathbf{y})$ for some $t > 0$.

In Lemma 8, we saw how expressing an estimator as the optimal value of a convex optimization problem allows us to take full advantage of convex duality in order to derive bounds on the estimator. We therefore start our analysis of (3.17) by characterizing its dual problem.

Lemma 16. *The dual problem of* (3.17) *is*

$$\max_{u \in \mathbb{R}^n} \left\{ \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \|\mathbf{y} - u\|_2^2\right) - \frac{1}{2\lambda}\left\|\frac{\mathbf{X}^T u}{n}\right\|_\infty^2 \right\}.$$

Similar arguments as in Lemma 8 give a bound expressing $\check{\sigma}_\lambda^2$'s closeness to the oracle estimator of σ^2.

Lemma 17. *If* $\lambda \geq n^{-1} \left\| \mathbf{X}^T(\varepsilon/\sigma) \right\|_\infty$, *then*

$$-2\lambda\sigma^2 \left(\frac{\|\beta^*\|_1}{\sigma} + \frac{1}{4} \right) \leq \check{\sigma}_\lambda^2 - \frac{1}{n} \|\varepsilon\|_2^2 \leq 2\lambda \|\beta^*\|_1^2 .$$

Comparing with Lemma 8, we see that the condition on λ depends only on a quantity $\varepsilon/\sigma \sim N(0, I_n)$ that is independent of σ^2. Indeed, this leads to a mean squared error bound with the desired property of λ not depending on σ.

Theorem 18. *Suppose that each column* $\mathbf{X}_j$ *of the matrix* $\mathbf{X} \in \mathbb{R}^{n \times p}$ *has been scaled so that* $\|\mathbf{X}_j\|_2^2 = n$ *for all* $j = 1, \ldots, p$, *and* $\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$. *Then, for any constant* $M > 1$, *the organic lasso estimator* (3.17) *with* $\lambda = (2Mn^{-1} \log p)^{1/2}$ *satisfies the following relative mean squared error bound:*

$$\mathrm{E}\left\{ \left(\frac{\check{\sigma}_\lambda^2}{\sigma^2} - 1 \right)^2 \right\} \leq \left\{ \left(8M + 8 \frac{p^{1-8M}}{\log p} \right)^{1/2} \max\left(\frac{\|\beta^*\|_1^2}{\sigma^2}, \frac{\|\beta^*\|_1}{\sigma} + \frac{1}{4} \right) \left(\frac{\log p}{n} \right)^{1/2} + \left(\frac{2}{n} \right)^{1/2} \right\}^2 . \tag{3.18}$$

Compared with Theorem 9, the organic lasso estimator of σ^2 retains the same rate in terms of n and p but has a slower rate in terms of $\sigma^{-1} \|\beta^*\|_1$. Importantly, though, the value of λ attaining (3.18) does not depend on σ. As in Remark 11, similar high-probability bounds can be obtained for ε with bounded polynomial moments.

Although not central to our main purpose, the organic lasso estimator (3.16) of β^* is interesting in its own right. The following theorem gives a slow rate bound in prediction error.

Theorem 19. *For any* $L > 0$, *the solution to* (3.16) *with* $\lambda = \{2n^{-1}(\log p + L)\}^{1/2}$ *has the following bound on the prediction error with probability greater than* $1 - e^{-L}$:

$$\frac{1}{n} \left\| \mathbf{X}\check{\beta}_\lambda - \mathbf{X}\beta^* \right\|_2^2 \leq \left(\sigma^2 + 4 \|\beta^*\|_1^2 \right) \left(\frac{2 \log p + 2L}{n} \right)^{1/2} .$$

In Appendix B.10, we provide mappings between the path of the natural lasso, $\{\hat{\beta}_\lambda : \lambda > 0\}$, and the path of the organic lasso $\{\check{\beta}_\lambda : \lambda > 0\}$. We also include a fast-rate prediction error bound of (3.16) under a compatibility condition.

3.5 Simulation studies

3.5.1 Simulation settings

Reid et al. (2016) carry out an extensive simulation study to compare many error variance estimators. We have matched their simulation settings fairly closely, so that the performance comparison with various other methods mentioned in Reid et al. (2016) can be inferred. Specifically, all simulations are run with $p = 500$ and $n = 100$. Each row of the design $\mathbf{X}$ is generated from a multivariate $N(\mathbf{0}, \Sigma)$, with $\Sigma_{ij} = \rho \in (0, 1)$ for $i \neq j$ and $\Sigma_{ii} = 1$. To generate β^*, we randomly select the indices of $\lceil n^\alpha \rceil$ (out of p) nonzero elements where $\alpha \in (0, 1)$, and each of the nonzero elements has a value that is randomly drawn from a Laplace distribution with rate 1. The error variance is generated using $\sigma^2 = \tau^{-1}\beta^{*T}\Sigma\beta^*$ for $\tau > 0$. Finally, $\mathbf{y}$ is generated following (3.1).

Each model is indexed by a triplet (ρ, α, τ), where ρ captures the correlation among features, α determines the sparsity of β^*, and τ characterizes the signal-to-noise ratio. We vary $\rho, \alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\tau \in \{0.3, 1, 3\}$. We compute a Monte Carlo estimate (based on 1000 replicates) of both the mean squared error $E\{(\hat{\sigma}/\sigma - 1)^2\}$ and $E(\hat{\sigma}/\sigma)$ as the measure of performance. The methods in comparison include: (a) the naive estimator (3.3) with $\hat{\beta}_\lambda$ in (3.2); (b) the degrees of freedom adjusted estimator $\hat{\sigma}^2_R$ in (3.5) (Reid et al. 2016); (c) the

square-root/scaled lasso (Belloni et al. 2011, Sun & Zhang 2013); (d) the natural lasso (3.7), and (e) the organic lasso (3.17). As a benchmark, we also include the oracle $n^{-1}\|\varepsilon\|_2^2$.

3.5.2 Methods with regularization parameter selected by cross-validation

We carry out two sets of simulations. In the first set, we compare the performance of the aforementioned methods with regularization parameter selected in a data-adaptive way. In particular, five-fold cross-validation is used to select the tuning parameter for each method.

Due to space constraints, we present a subset of the results in Fig 3.1 (with additional results presented in Appendix B.12). The result for the square-root/scaled lasso is averaged over 100 repetitions due to the large computational time. For all other methods, the results are averaged over 1000 repetitions. Overall, the natural lasso does well in adjusting the downward bias of the naive estimator, while other methods tend to produce under-estimates. In each panel, we fix signal-to-noise ratio (τ) and correlations among features (ρ), and vary model sparsity (α). All estimates get worse with growing α, except for the natural lasso, which improves as the true β^* gets denser. In particular, both the natural lasso and the organic lasso gain performance advantage over other methods when the underlying models do not satisfy conditions for the support recovery of the lasso solution. From left to right, Fig 3.1 illustrates the effect of increasing ρ. As observed in Reid et al. (2016), high correlations can be helpful: All curves approach the oracle as ρ increases. Finally, we find that the organic

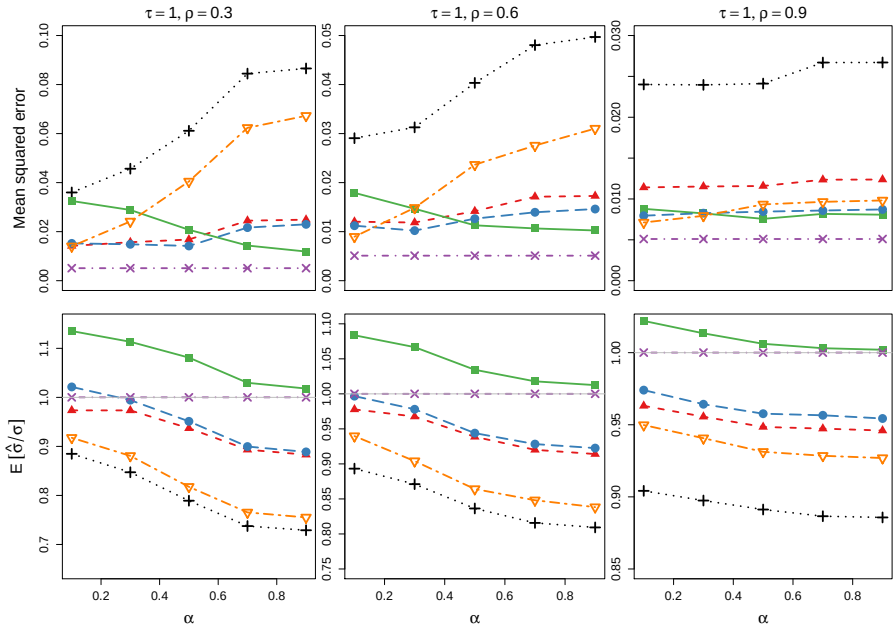

Figure 3.1: Simulation results of methods using cross-validation. From left to right, columns show the average (over 1000 repetitions) of the mean squared error (top panel) and $E(\hat{\sigma}/\sigma)$ (bottom panel) of various methods in three simulation settings. Line styles and their corresponding methods: + for naive, ▲ for $\hat{\sigma}_R^2$, ▽ for the square-root/scaled lasso, — for the natural lasso, ● for the organic lasso, × for the oracle.

lasso is uniformly better or equivalent to $\hat{\sigma}_R^2$.

Paired t-tests and Wilcoxon signed-rank tests show that the differences in mean squared errors of different methods are significant at the 5% level for almost all points shown in Fig 3.1.

Results in Appendix B.12 also show the natural lasso estimator doing well when the signal-to-noise ratio is low: the performance of all methods degrade as τ gets large. This is expected from Theorem 9 and Theorem 18, and is also observed in Reid et al. (2016).

3.5.3 Methods with fixed choice of regularization parameter

Although solving (3.17) is fast enough for one to use cross-validation with the organic lasso, Theorem 18 implies that $\lambda_0 = (2n^{-1} \log p)^{1/2}$ is a theoretically sound choice of regularization parameter. We also conjecture that a sharper rate may be obtainable at $\lambda_1 \geq \|\mathbf{X}^T\epsilon\|_\infty^2/n^2$, where $\epsilon \sim N(0,1)$. With high probability, $\|\mathbf{X}^T\epsilon\|_\infty^2/n^2 \approx \log(p)/n$. Thus, we also show the performance of the organic lasso with tuning parameter values equal to $\lambda_2 = \log(p)/n$, and λ_3, which is a Monte Carlo estimate of $\|\mathbf{X}^T\epsilon\|_\infty^2/n^2$.

We compare the organic lasso at these three fixed values of tuning parameter to the square-root/scaled lasso estimator (3.11) of error variance, which is another method whose theoretical choice of λ does not depend on σ. Sun & Zhang (2012) find that λ_0 works very well for (3.11), which we denote by scaled(1), and Sun & Zhang (2013) propose a refined choice of λ, which is proved to attain a sharper rate, denoted by scaled(2).

Fig 3.2 shows similar patterns as Fig 3.1. Specifically, large value of ρ helps all methods, while performance generally degrades for denser β^*. Although not shown here, all methods struggle as τ increases. The organic lasso with λ_0 performs poorly, while the organic lasso with λ_2 and λ_3 do quite well, generally outperforming the square-root/scaled lasso based methods.

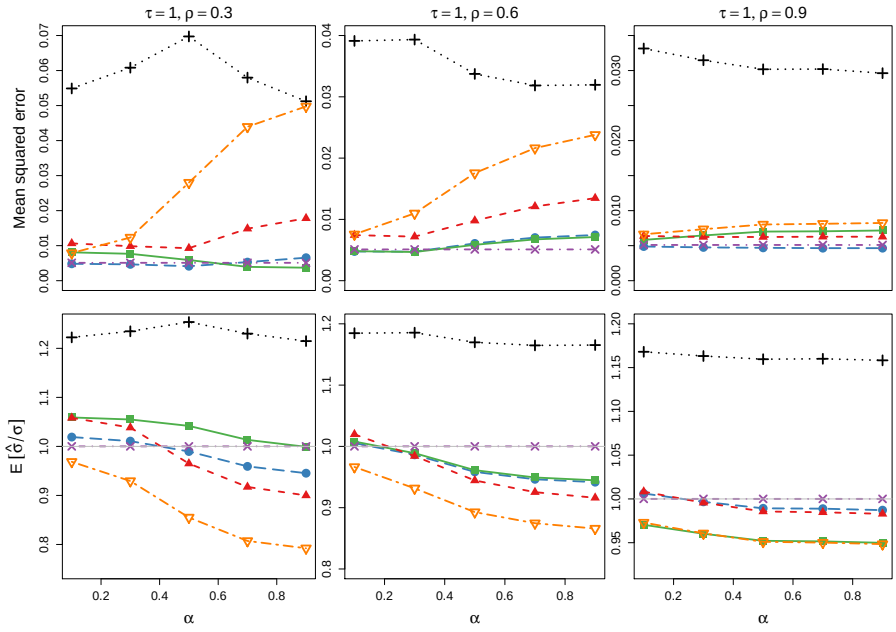

Figure 3.2: Simulation results of methods using pre-specified regularization parameter values. From left to right, column show the average (over 1000 repetitions) of the mean squared error (top panel) and $E(\hat{\sigma}/\sigma)$ (bottom panel) of various methods in three simulation settings. Line styles and their corresponding methods: for organic (λ_0), for organic (λ_2), for organic (λ_3), for scaled(1), for scaled (2), for the oracle.

3.6 Error estimation for Million Song dataset

We apply our error variance estimators to the Million Song dataset.[1] The data consist of information about 463715 songs, and the primary goal is to model the release year of a song using $p = 90$ of its timbre features. The dataset has a very large sample size so that we can reliably estimate the ground truth of the target of estimation on a very large set of held out data. In particular, we randomly select half of the songs for this purpose and use $\bar{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}_{LS}\|_2^2/(n-p)$ to form our ground truth, where $\hat{\beta}_{LS}$ is the least-squares estimator of β^*. In practice,

[1]The whole data set can be obtained at https://labrosa.ee.columbia.edu/millionsong/. We consider a subset of the whole data, which is available at https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd.

model (3.1) may rarely hold, which alters the interpretation of error variance estimation. Suppose the response vector y has mean μ and covariance matrix Σ. Then $\bar{\sigma}^2$ can be thought of as an estimator of the population quantity

$$\min_{\beta} \frac{1}{n} E\left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2\right) = \frac{1}{n}\operatorname{tr}(\Sigma) + \frac{1}{n}\left\|\left(I - XX^{+}\right)\mu\right\|_2^2.$$

In the special case where $\Sigma = \sigma^2 I_n$ and $\mu = X\beta^*$, as in (3.1), then $\bar{\sigma}^2$ reduces to the linear model noise variance σ^2.

From the remaining data that was not previously used to yield $\bar{\sigma}^2$, we randomly form training datasets of size n and compare the performance of various error variance estimators. We vary n in $\{20, 40, 60, 80, 100, 120\}$ to gauge the performance of these methods in situations in which $n < p$ and $n \approx p$. For each n, we repeat the data selection and error variance estimation on 1000 disjoint training sets, and report estimates of the mean squared error $E\{(\hat{\sigma}/\bar{\sigma} - 1)^2\}$ in Table 3.1 and estimates of $E(\hat{\sigma}/\bar{\sigma})$ in Appendix B.12.

All methods produce a substantial performance improvement over the naive estimator for a wide range of values of n. The natural and organic lassos with cross validation perform either better or comparably to $\hat{\sigma}_R^2$ and are in some (but not all) cases outperformed by scaled(2). The natural lasso shows some upward bias (which as we noted before is less problematic than downward bias) when n gets large. The organic lasso with the fixed choices λ_2 or λ_3 perform extremely well for all n.

Future research directions include the analysis of the proposed methods with smaller values of λ, and extending the natural parameterization to penalized non-parametric regression. Finally, an `R` (R Core Team 2017) package, named `natural` (Yu 2017), is available on the Comprehensive R Archive Network, implementing our estimators.

Table 3.1: Mean squared error of noise variance estimation for Million Song dataset

n	20	40	60	80	100	120
naive	17.02 (0.68)	8.48 (0.41)	5.28 (0.26)	3.80 (0.17)	3.03 (0.13)	2.43 (0.10)
$\hat{\sigma}^2_R$	10.74 (0.45)	5.92 (0.29)	3.57 (0.17)	2.57 (0.11)	2.23 (0.10)	1.75 (0.08)
natural	8.82 (0.38)	5.23 (0.27)	3.47 (0.16)	2.61 (0.12)	2.39 (0.11)	2.01 (0.09)
organic	8.08 (0.32)	4.23 (0.20)	2.59 (0.12)	2.00 (0.08)	1.72 (0.08)	1.54 (0.07)
scaled(1)	7.43 (0.37)	4.92 (0.25)	3.84 (0.17)	3.08 (0.13)	2.94 (0.12)	2.75 (0.11)
scaled(2)	7.11 (0.28)	3.36 (0.15)	2.23 (0.10)	2.57 (0.83)	1.61 (0.07)	1.46 (0.07)
organic(λ_2)	5.87 (0.24)	3.17 (0.14)	1.93 (0.09)	1.40 (0.06)	1.20 (0.05)	1.02 (0.05)
organic(λ_3)	5.72 (0.24)	3.15 (0.14)	1.99 (0.09)	1.45 (0.07)	1.28 (0.05)	1.12 (0.05)

Mean and standard errors (over 1000 replications) of the squared error of various methods. Each entry is multiplied by 100 to convey information more compactly.

Acknowledgement

We thank Irina Gaynanova for a useful conversation that helped us prove Propositions 12 and 13. JB was supported by an NSF CAREER grant, DMS-1653017.

CHAPTER 4

RELUCTANT INTERACTION MODELING

Material of this chapter is not published yet. It is a joint work with Ryan Tibshirani and Jacob Bien.

4.1 Introduction

Given a response variable and several features of interest, it is a fundamental problem in many fields to identify which features are relevant for predicting the response. This problem becomes a major statistical challenge when the number of features collected exceeds the sample size, a situation that has become increasingly common in contemporary scientific research (in fields such as genetics, medicine, and the social sciences). The last two decades have witnessed many advances in addressing this "high-dimensional" challenge (Tibshirani 1996, Fan & Li 2001, Zou & Hastie 2005, Candes & Tao 2007, Fan & Lv 2008, Sun & Zhang 2012, Belloni et al. 2011, etc.), and the computational and theoretical properties of these methods have been well studied.

However, in many situations, modeling the response as a linear function of the features (i.e., as *main effects*) might not be sufficient to characterize the full complexity of the relationship. In many settings, one finds that interactions between features account for variability in the response that cannot be explained by an additive function of the features alone. It is plausible that many biological phenomena, e.g., effects of various behaviors, exposures, and genetic factors on disease rates are not additive. Indeed, in genome-wide association studies

(GWAS), the interaction effects among single-nucleotide polymorphisms (SNPs) and their interactions with other genetic or environmental factors have been found to be critical in understanding how certain human diseases formulate (Cordell 2009). Moreover, many problems in traditional statistics, including experimental design and nonlinear regression, naturally involve interaction effects.

We consider the following two-way interaction model:

$$Y = \sum_{j=1}^{p} X_j \beta_j^* + \sum_{j \le k} X_j X_k \gamma_{jk}^* + \varepsilon = X^T \beta^* + Z^T \gamma^* + \varepsilon, \tag{4.1}$$

where $X \in \mathbb{R}^p$ is a p-dimensional random vector of main effects, $Z = (X_1 * X_1, X_1 * X_2, \ldots, X_p * X_p) \in \mathbb{R}^{(p^2+p)/2}$ is the random vector of all pairwise interactions of X, and ε is an additive zero-mean noise independent of X. Model (4.1) extends a typical linear model (in main effects X), and γ^* characterizes how the pairwise interactions among features relate to the response. Although our method could be easily generalized to modeling interactions of higher order, for simplicity we restrict ourselves to the two-way interaction model (4.1).

With the sparsity assumptions that only a small number of components in β^* and γ^* are nonzero, one might naturally consider solving a lasso (Tibshirani 1996) using all the main effects and the interactions. We will call this method the *all pairs lasso* (APL). Theoretically, under the standard compatibility conditions, APL's prediction mean squared error rate is of order $O(\log(p^2)/n)$ in n and p, and is identical in rate to the *main effects lasso* (MEL), which is $O(\log(p)/n)$. However, the practical computational difference between APL and MEL can be dramatic, especially for large values of p. In practice, APL quickly becomes infeasible to compute as p gets large. Standard lasso solvers require passing the whole augmented design matrix of main effects and interactions, which requires $O(np^2)$

space. Moreover, even if we compute the interactions on the fly when solving the lasso, the state-of-art coordinate descent type of algorithm requires multiple passes over all $O(p^2)$ variables until convergence.

Motivated by these observations, we introduce in this paper a computationally viable approach to interaction modeling, called *sprinter* (for sparse regularized interaction modeling). In particular, our contribution includes the following:

- We propose a new principle in large-scale interaction modeling, which says that one should prefer main effects over interactions given similar prediction performance. We emphasize that this principle is distinct from (although reminiscent of) the common heredity principle. Sprinter is a multiple-stage method that honors this new principle: in the first stage it tries to capture as much of the variability in the response as possible without resorting to interactions; in the second stage it includes only interactions that capture signal that cannot be captured by main effects. In this sense, sprinter is a "reluctant" interaction selection procedure.
- By adhering to this principle, sprinter allows for interaction modeling on unprecedented problem sizes (for a method not relying on the heredity principle) without compromising practical statistical performance. In particular, sprinter attains empirical statistical performance which is competitive with APL while being much easier to compute (both in terms of time and storage): sprinter fits an interaction model with 2000 main effects about 100 times faster than APL, and it fits a problem with 140000 main effects (about 10 billion interactions) with 5-fold cross-validation in under 7 hours on a single CPU. In addition, sprinter achieves statistical

performance that compares favorably with alternative methods that are also computationally efficient.

- Theoretically, we show that the prediction error rate of sprinter is comparable to APL while being much more computationally efficient.

4.1.1 Related methods

Variable selection in large-scale interaction models (i.e., when p is large) is computationally very difficult, as the number of interactions, i.e., $\binom{p}{2}$, grows quadratically with p. Assumptions on the interaction structure are usually made to facilitate computation and interpretation. Hierarchy (Peixoto 1987), also known as heredity (Hamada & Wu 1992) or marginality (Nelder 1977), is a standard assumption that greatly improves computational feasibility. The hierarchical assumption is that an interaction effect is in the model only if either (or both) of the main effects corresponding to the interaction are in the model.

Various methods using the hierarchical assumption have been proposed, which can be broadly categorized into single stage methods and multiple stage methods. A single stage method essentially formulates the problem of estimating (β^*, γ^*) in (4.1) as an optimization problem which minimizes a penalized loss criterion jointly with respect to β^* and γ^*. The penalty is designed specifically so that the resulting estimate honors certain hierarchical assumptions. While this type of method enjoys good theoretical analysis, it is only computationally valid for moderate problem size (Efron et al. 2004, Turlach 2004, Zhao et al. 2009, Yuan et al. 2009, Choi et al. 2010, Radchenko & James 2010, Schmidt & Murphy 2010, Bien et al. 2013, Haris et al. 2016, Lim & Hastie 2015, Haris et al. 2016, She

et al. 2018).

The other type of method estimates the unknown parameters in multiple stages. Usually in the first stage the vector of main effects β^* is estimated. The hierarchical assumption then implies that one can constrain the search space of potential interactions to only those interactions that consist of main effects selected in the first stage. Therefore, multiple stage methods usually enjoy superior computational efficiency (Wu et al. 2009, 2010, Hao et al. 2018). However, the theoretical guarantees of the resulting estimator are harder to achieve because the analysis hinges on not missing any nonzero elements of β^*, which usually requires stronger assumptions (e.g., Zhao & Yu 2006, Hao et al. 2018). Hao & Zhang (2014) propose two forward-selection-based greedy algorithms to select interactions under the hierarchical assumption. They show that their proposed methods enjoy screening consistency under certain assumptions. Shah (2016) propose a general framework (for various penalty functions) that builds a set of active interactions iteratively, guided by the idea of hierarchy. The method attains computational efficiency by utilizing the results from previous computation and enjoys screening properties under certain assumptions.

However, the hierarchical assumption need not always hold. For example, Culverhouse et al. (2002) study purely epistatic models (i.e., those with no main effects). When the underlying interaction structure is not hierarchical, all the aforementioned methods based on the hierarchy assumptions can suffer.

Fan et al. (2016) propose the *interaction pursuit* (IP), a novel two-stage method to detect interaction effects without the hierarchical assumptions. The proposed method is computationally efficient because it does not directly screen the $O(p^2)$ interactions, but instead it focuses on identifying the set of "active interaction

variables" $\mathcal{A}$, i.e., a subset of the $O(p)$ main effects that are involved in the nonzero interactions. In the second stage, the method only considers interactions constructed from the recovered interaction variables. This method is efficient and can be quite effective. Its success relies on proper recovery of $\mathcal{A}$. When the interactions are concentrated among a small set $\mathcal{A}$ of interaction variables, $\mathcal{A}$ can be easily recovered. The most challenging situation for this method is when there is no such concentration of interactions over a small set of main effects. Thanei et al. (2016) consider a randomized algorithm that solves each step of APL approximately by solving a closest-pair problem. By doing so, they show that the computational complexity of their method is sub-quadratic in p.

4.1.2 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we propose our method, which honors a new principle in large-scale interaction modeling. In particular, our method prefers using main effects over interactions in fitting the response, and an interaction is only included in the model if no main effects is helpful in explaining response. We give an example that motivates this principle, and a re-parameterized model to rigorously characterize its feasibility. In Section 3, we formally introduce *sprinter*, and discuss its practical implementation as well as its computational complexity. The theoretical analysis of sprinter is then given in Section 4, where we present, under certain conditions, the prediction error rates in n and p. In Section 5 we study the empirical performance of sprinter in various simulation studies and in a data example respectively.

4.1.3 Notation

Let $q = (p^2 + p)/2$, and denote $\Sigma = \text{Cov}(X) \in \mathbb{R}^{p\times p}$, $\Psi = \text{Cov}(Z) \in \mathbb{R}^{q\times q}$, and $\Phi = \text{Cov}(X, Z) \in \mathbb{R}^{p\times q}$. Given a matrix $M \in \mathbb{R}^{a\times b}$ and an index set T, denote M_T as the $a \times |T|$ sub-matrix of M with columns selected from T. On a sample level, we denote $\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_p\right) \in \mathbb{R}^{n\times p}$ as the design matrix with each column $\mathbf{X}_j \in \mathbb{R}^n$ consisting of all observations of variable X_j (for $j = 1, \ldots, p$). Similarly $\mathbf{Z} \in \mathbb{R}^{n\times q}$ is the sample matrix of $Z = (X_1 * X_1, X_1 * X_2, \ldots, X_p * X_p)$, $\mathbf{y} \in \mathbb{R}^n$ is the response vector, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of n independent samples of random error ε. We denote $\mathbf{X}_j * \mathbf{X}_k$ to be the element-wise product of $\mathbf{X}_j$ and $\mathbf{X}_k$ and $\mathbf{X}_j^2 = \mathbf{X}_j * \mathbf{X}_j$. We let $\widehat{\text{corr}}(\mathbf{X}_j, \mathbf{X}_k)$ stand for the sample correlation between $\mathbf{X}_j$ and $\mathbf{X}_k$, and let $\widehat{\text{sd}}(\mathbf{X}_j)$ to be the sample standard deviation of $\mathbf{X}_j$. Finally, $[p]$ is the set $\{1, 2, 3 \ldots, p\}$.

4.2 A new principle in large-scale interaction modeling

One main reason for the heavy computational burden of APL is that multiple passes over $O(p^2)$ variables (including both main effects and interactions) are required until convergence. What if we could afford only a single pass over all $O(p^2)$ variables and are not able to have $O(p^2)$ memory? In this case, one might consider a simple screening procedure that screens out irrelevant variables based on some measure of importance.

Yet this idea treats the p main effects and the $\binom{p}{2}$ interactions equivalently. The basic premise of our method is that we would like to fit the response as well as possible using only the main effects (or more generally a set of $O(p)$ features based on main effects) and then only include interaction terms for what cannot

be captured by main effects.

For example, consider a simple model $Y = X_1 + X_2 + X_1 * X_2$, where $X_1 = \mathbb{1}_A$ and $X_2 = \mathbb{1}_B$ are the indicator variables of events A and B, respectively. Suppose further that with high probability $A \subseteq B$, so that $X_1 * X_2 = \mathbb{1}_A * \mathbb{1}_B = \mathbb{1}_{A \cap B} \approx \mathbb{1}_A$; hence, the main effect $X_1 = \mathbb{1}_A$ can be used in place of the interaction $X_1 * X_2$, i.e., $Y \approx 2X_1 + X_2$. This main effects only explanation of Y is simpler to understand and yet explains Y nearly as well as the original model with the interaction term.

As a second example, suppose two main effects X_j and X_k are highly correlated. In such a case, the interaction $X_j * X_k$ is then not much different from the squared effect X_j^2 (or X_k^2). Thus when main effects are highly correlated, much of the interaction signal can be captured using only the p squared main effect terms. In more general scenarios, specific interactions may be strongly correlated with linear combinations of main effects and (or) squared effects, which means that we could get equivalently predictive models without using that interaction.

These examples demonstrate that main effects, or simple functions of main effects, are able in some cases to act as useful handles in approximating interactions. This observation leads us to propose a new principle in large-scale interaction modeling:

> **The reluctant interactions selection principle (RISP):** *One should prefer a main effect over an interaction if all is equal.*

Leaning on main effects more heavily than interactions is advantageous for two reasons. First, main effects are easier to interpret than interactions, thus when

presented with two models that predict the response equivalently, we should favor the one that relies on fewer interactions. But, second, we will show in this paper that prioritizing main effects (or simple functions of main effects such as squared terms) can lead to great computational savings (both in terms of time and memory). The key reason for these savings is that when p is large, the total number of main effects is far smaller than the number of interactions.

We emphasize that the proposed principle is different from the well-known hierarchy assumption. While both principles simplify the search of interactions by focusing on certain main effects, our principle does not explicitly tie an interaction to its corresponding main effects. For example, an interaction $X_3 * X_4$ could be highly correlated with a linear combination of X_1 and X_2, which may lead us to exclude $X_3 * X_4$. This logic is very different from the logic used in the hierarchical principle.

More specifically, model (4.1) expresses the signal in terms of a main effects signal term, $X^T\beta^*$, and an interactions signal term, $Z^T\gamma^*$. If X and Z were uncorrelated, this would be a unique decomposition. However, as demonstrated in the examples, there can be "overlap". Let $X^T\vartheta^*$ be the part of $Z^T\gamma^*$ that can be explained by a linear combination of X, i.e.,

$$\vartheta^* := \underset{\vartheta\in\mathbb{R}^p}{\arg\min}\, \mathrm{Var}\left(Z^T\gamma^* - X^T\vartheta\right) = \mathrm{Cov}\,(X)^{-1}\,\mathrm{Cov}\,(X,Z)\,\gamma^* = \Sigma^{-1}\Phi\gamma^*. \tag{4.2}$$

We can then write (4.1) as

$$Y = X^T\left(\beta^* + \vartheta^*\right) + W^T\gamma^* + \varepsilon, \tag{4.3}$$

where

$$W := Z - \Phi^T\Sigma^{-1}X$$

is the "pure" interaction effects that cannot be captured by linear combinations of X, and it holds that $\mathrm{Cov}(X, W) = 0$. We denote the covariance of the pure interactions

$$\Omega := \mathrm{Cov}(W) = \mathrm{Cov}(Z, W) = \mathrm{Cov}(Z) - \mathrm{Cov}(X)\Sigma^{-1}\Phi = \Psi - \Phi^T\Sigma^{-1}\Phi. \tag{4.4}$$

By fitting Y using only linear combinations of X, we ignore the effect from W. We will see in the following theoretical results that the zero covariance between X and W ensures that the model misspecification when ignoring W has minimal effects in subsequent procedures. Actually, this simplicity of theoretical analysis from the "orthogonality" between main effects and interactions is also observed in Hao & Zhang (2014), where X is assumed to follow a symmetric distribution with respect to $\mathbf{0}$. In such a case, we have that $\Phi = \mathrm{Cov}(X, Z) = \mathbf{0}$, which implies that $\vartheta^* = \mathbf{0}$ and $W = Z$. Our method does not require the symmetry of the distribution of main effects, and thus allows for more general covariance structure between main effects X and interactions Z.

Finally, we note that X in (4.3) can be generalized to be a random vector containing main effects and simple functions of main effects, i.e., the squared effects, spline functions of main effects, or even regression trees based on main effects. For example, when X is Gaussian, main effects and interactions are known to be uncorrelated; however, when squared main effects are added to X, then we no longer have $Z = W$.

4.3 Main proposal: sprinter

In this section, we describe a new method, called sprinter, that is based on the RISP principle. The proposed method has three steps:

- In Step 1, we fit the response as well as possible using only the $O(p)$ main effects variables (or simple univariate functions of these). This step purposely gives preference to main effects, corresponding to the RISP principle described in Section 4.2.
- In Step 2, we perform a single pass over all interactions to identify interaction signal that was not captured in Step 1. Because each of the $O(p^2)$ interactions is only computed and used once, this step requires far less time and memory than APL, which requires passing over all interactions multiple times.
- In Step 3, we fit a lasso (or any other user-specified variable selection method) on all main effects and the *pure interactions* that were selected in Step 2. "Pure" here is in the sense of W introduced in the previous section. By doing so, the total number of variables is $O(p)$ and thus is far more efficient to compute than APL.

Algorithm 4: sprinter

Require: Main effects $\mathbf{X} \in \mathbb{R}^{n \times p}$, response $\mathbf{y} \in \mathbb{R}^n$, $\eta > 0$

1: **Step 1**:

2: Fit a regularized regression model of the response $\mathbf{y}$ on $\mathbf{X}$.

3: Compute the residual $\mathbf{r}$.

4: **Step 2**:

5: For a tuning parameter η, screen based on the residual:

$$\hat{\mathcal{I}}_\eta = \left\{ j \in [q] : \widehat{\mathrm{sd}}(\mathbf{r}) |\widehat{\mathrm{corr}}\left(\mathbf{Z}_j, \mathbf{r}\right)| > \eta \right\},$$

6: **Step 3**:

7: Fit a lasso of the response $\mathbf{y}$ on $\mathbf{X}$ and all the pure interactions in $\hat{\mathcal{I}}_\eta$.

In Step 1, $\mathbf{X}$ could be replaced by any design matrix of $O(p)$ main effects related variables (in such a case, Step 2 would still only consider interactions between the original p main effects). Step 2 can be considered as a *sure independence screening* (SIS; Fan & Lv 2008, Barut et al. 2016) of all interactions using the residual from Step 1. In practice, the optimal value of the tuning parameter η is usually unknown and thus requires tuning. Instead, we consider screening using

$$\hat{\mathcal{I}}_m = \left\{ j \in [q] : |\widehat{\text{corr}}\left(\mathbf{Z}_j, \mathbf{r}\right) \text{ is among the } m \text{ largest} \right\}.$$

This top-m approach is standard in screening based variable selection methods (Fan & Lv 2008, Barut et al. 2016) and large-scale interaction modeling approaches (Fan et al. 2016, Niu et al. 2018). Popular choices of values of m include n and $\lceil n/\log(n) \rceil$.

These first two steps are built around the RISP principle. Given a set of highly correlated variables, the lasso tends to select just one of them. Thus, if an interaction is highly correlated with one or more main effects, APL may very well select the interaction. By contrast, sprinter explicitly prioritizes the main effects (in Step 1). The interaction will only be selected (in Step 2) if it can capture something in the signal that the main effects cannot.

In Step 3, one would like to use the "pure interactions" $\mathbf{W}_{\hat{\mathcal{I}}_\eta}$ (as opposed to the actual interactions $\mathbf{Z}_{\hat{\mathcal{I}}_\eta}$) in fitting the final model. In the setting where Σ and Φ are known, or in the semi-supervised learning setting where one has a vast number of training samples of X and thus could obtain accurate estimates of Σ and Φ, the matrix of pure interactions $\mathbf{W}$ is obtainable. In other settings where $\mathbf{W}$ is not directly available, we replace $\mathbf{W}$ with some approximation $\hat{\mathbf{W}}$. Step 3 does not specify the method to get $\hat{\mathbf{W}}$, and as we will see in Corollary 23, Step 3's

success depends only on $\hat{\mathbf{W}}$ being a decent approximation to $\mathbf{W}$. As an example, one could simply set $\hat{\mathbf{W}}$ with each column $\hat{\mathbf{W}}_j = \mathbf{Z}_j - \mathbf{X}\hat{\phi}_j$ for $j \in \hat{I}_\eta$, and $\hat{\phi}_j$ is a solution to the following lasso problem

$$\hat{\phi}_j \in \arg\min_{\phi \in \mathbb{R}^n} \left(\frac{1}{n} \left\| \mathbf{Z}_j - \mathbf{X}\phi \right\|_2^2 + \nu \left\| \phi \right\|_1 \right). \tag{4.5}$$

Later in Corollary 24 we show the prediction error rate of Step 3 when $\hat{\mathbf{W}}$ is calculated as above. For $\mathbf{Z}_j = \mathbf{X}_k * \mathbf{X}_\ell$, a computationally more efficient alternative of getting $\hat{\mathbf{W}}_j$ is to get the residual from a least squares regression of $\mathbf{Z}_j$ on $\mathbf{X}_k$ and $\mathbf{X}_\ell$. In general, we find very little difference in the effect of using different $\hat{\mathbf{W}}$. For example, we find that simply using $\hat{\mathbf{W}} = \mathbf{Z}$ still generally gives favorable performance.

4.3.1 Computation

With a value of $m \leq n$, the required computation in both Step 1 and Step 3 are no worse than fitting a lasso with $p + n$ features. If computing $\hat{\mathbf{W}}_{\hat{I}_m}$ is needed, and we use, for example, (4.5), then solving an extra m lasso problems with p features would be necessary. However, the major computational burden lies in Step 2, where $O(p^2)$ sample correlations are computed. It is thus essential for this step to be implemented as efficiently as possible, both in terms of computational time and storage.

We compute the sample correlation between each interaction and the residual from Step 1 on the fly. In the meantime, a min-heap is maintained to keep the index pairs of the interactions that attain the m largest sample correlations. This ensures that we won't have to store $O(p^2)$ elements. The time complexity of Step 2 is thus $O(np^2 + p^2 \log m)$, where $O(p^2 \log m)$ is for main-

taining the m largest elements in a data stream of size $O(p^2)$ by maintaining a min-heap of size m (Cormen et al. 2009). Thus the total time complexity is $O(np^2) + O(p^2 \log m) + O(n(p+m)\min\{n, p+m\})$. Note that the whole algorithm only requires $O(n(p+m))$ storage.

4.4 Theoretical analysis

The theoretical analysis is based on the following assumptions:

A1 Distribution: $X = (X_1, \ldots, X_p)$ follows a sub-Gaussian distribution with covariance matrix Σ, and ε is zero-mean sub-Gaussian noise independent of X with $\mathrm{Var}(\varepsilon) = \sigma^2$.

A2 Dimensionality: There exist some constants A_1 and A_2 such that $\|\gamma^*\|_1 \leq A_1 n^{\xi_1}$, $\log(q) \leq A_2 n^{\xi}$ for some constants $\xi, \xi_1 > 0$.

A3 The prediction $\mathbf{X}\check{\theta}$ from Step 1 has the following error rate with high probability

$$\frac{1}{\sqrt{n}} \left\| \mathbf{X}\theta^* - \mathbf{X}\check{\theta} \right\|_2 \leq n^{-\tau}. \tag{4.6}$$

A4 Minimum signal strength: Let

$$\mathcal{I} = \left\{ j \in [q] : \mathrm{E}(W_j^2)\gamma_j^{*2} > \alpha \right\}, \tag{4.7}$$

for some $\alpha \geq 0$. Then for some $\kappa \geq 0$ and some constant $C_\kappa > 0$, we have

$$\min_{j \in \mathcal{I}} \mathrm{Cov}(Z_j, W^T\gamma^*) = \min_{j \in \mathcal{I}} \Omega_j^T \gamma^* \geq \frac{3C_\kappa}{2} n^{-\kappa}.$$

Assumptions similar to **A1** and **A2** have also been made in previous work that has involved the screening property for high-dimensional interaction modeling (see, e.g., Hao & Zhang 2014, Fan et al. 2016). In particular, **A1** is a very

general distributional assumption on the random design; unlike other methods, we neither require that the distribution of X is symmetric nor has mean zero. **A2** specifies the underlying problem dimensions: we allow the total number of interactions q to grow exponentially with the sample size n, and $\|\gamma^*\|_1$ to grow as a polynomial function of n;

A3 requires that Step 1 attains a sufficiently good prediction error rate. Recall that Step 1 fits a misspecified model since it ignores the pure interaction signal $W^T\gamma^*$ in (4.3). One might ask if Assumption **A3** can actually be met. Section 4.4.1 provides an affirmative answer to this question, showing that (for certain values of τ) a lasso in Step 1 can for example be used.

Step 2 is a screening process based on the sample correlation with the residual:

$$\hat{\mathcal{I}}_\eta = \left\{ j \in [q] : \widehat{\mathrm{sd}}(\mathbf{r}) |\widehat{\mathrm{corr}}\left(\mathbf{Z}_j, \mathbf{r}\right)| > \eta \right\},$$

where η is a tuning parameter, and $\mathbf{r}$ is the residual from Step 1. The hope is that $\hat{\mathcal{I}}_\eta$ has the screening property, i.e., that it retains all the important pure interactions, where "important" is defined as being in $\mathcal{I}$: in particular, according to (4.7) we would consider an interaction important if the marginal pure interaction effect $\mathrm{E}\{(W_j\gamma_j^*)^2\}$ is sufficiently large. Recall that $W^T\gamma^*$ is the part of the interaction signal that cannot be explained by linear combinations of main effects. Thus, intuitively $\mathcal{I}$ is the set of interaction indices accounting for most of this pure interaction signal. When $\alpha = 0$, we get $\mathcal{I} \subseteq \mathrm{supp}(\gamma^*)$. In typical interaction modeling, the goal would be to recover $\mathrm{supp}(\gamma^*)$; however, in our reluctant interaction selection framework, we do not care about recovering an interaction $j \in \mathrm{supp}(\gamma^*)$ if $\mathrm{E}[W_j^2] = 0$. That is, if Z_j can be perfectly explained by a linear combination of main effects, then we do not care that $\gamma_j^* \neq 0$.

If Step 1 does a good job in capturing all signal from main effects, i.e., $\mathbf{X}\theta^*$, then $\mathbf{r}$ should essentially be the pure interaction signal $\mathbf{W}\gamma^*$ (with noise). For the screening property to hold, we require **A4**, which is similar to a "β-min" condition in the screening and selection consistency literature (see, e.g., Fan & Lv 2008, Raskutti et al. 2011). **A4** requires that the covariance between each actual interaction $\mathbf{Z}_j$ for $j \in \mathcal{I}$ and the pure interaction signal $W^T\gamma^*$ should not decay too fast to be detected.

However, by trivially taking $\eta = 0$, one can easily achieve the goal of not missing any important interactions, i.e., $\hat{\mathcal{I}}_\eta = [q]$. By doing so, Step 2 enjoys the screening property, yet there is no computational gain over APL. Therefore, we should also require that in addition to the property $\mathcal{I} \subseteq \hat{\mathcal{I}}_\eta$, Step 2 does not over-select many spurious interactions. The following theorem shows that Step 2 meets these two criteria.

Theorem 20 (Screening property in Step 2)**.** *Under Assumptions **A1** to **A4**, if* $\sqrt{2\max_j \Psi_{jj}}n^{-\tau} \le 6^{-1}C_\kappa n^{-\kappa}$ *and* $\xi + 2\xi_1 + 2\kappa < \frac{1}{2}$*, then with* $\eta = C_\kappa n^{-\kappa}$ *it holds that*

$$\mathcal{I} \subseteq \hat{\mathcal{I}}_\eta \quad \textit{and} \quad |\hat{\mathcal{I}}_\eta| \le 4C_\kappa^{-2}\lambda_{\max}\left(\mathrm{diag}(\Psi)^{-1/2}\Omega\right)\mathrm{Var}(Y)n^{2\kappa} \tag{4.8}$$

with probability greater than $1 - c_1\exp(-c_2 n^\xi)$ *for some constants* $c_1, c_2 > 0$*,*

Proof. See Appendix C.1. □

The optimal tuning parameter η in the screening step depends on the unknown quantities C_κ and κ. In practice, we adapt the same strategy as Fan & Lv (2008) and Fan et al. (2016), i.e., we take the m largest $|\hat{\omega}_j|$. Some popular choices of value of m include n and $\lceil n/\log(n)\rceil$. Under the assumptions that $\mathrm{Var}(Y) = O(1)$ and $\lambda_{\max}\left(\mathrm{diag}(\Psi)^{-1/2}\Omega\right) = O(n^{\xi_\lambda})$ for some some $\xi_\lambda > 0$, both of which are standard

assumptions (see, e.g., Fan & Lv 2008, Hao & Zhang 2014, Fan et al. 2016), Theorem 20 shows that the total number of interactions retained after the screening is a polynomial function of n. Section 4.4.2 relates Theorem 20 to previous work in high-dimensional interaction screening. From now on, we will suppress the notation dependence of $\hat{\mathcal{I}}_\eta$ on η for simplicity.

We now present the main theoretical result, which is a deterministic bound on the prediction error of Step 3. Recall that in Step 3 we solve the problem,

$$\left(\hat{\theta}, \hat{\gamma}\right) \in \underset{\theta \in \mathbb{R}^p, \gamma \in \mathbb{R}^{|\hat{\mathcal{I}}|}}{\arg\min} \left\{ \frac{1}{2n} \left\| \mathbf{y} - \mathbf{X}\theta - \hat{\mathbf{W}}_{\hat{\mathcal{I}}} \gamma \right\|_2^2 + \lambda \left(\|\theta\|_1 + \|\gamma\|_1 \right) \right\},$$

where $\hat{\mathbf{W}}$ is the design matrix of some approximated pure interaction variables.

Theorem 21 (Prediction error in the final step)**.** *Take*

$$\lambda \geq \max\left(\frac{1}{n} \max_{1 \leq j \leq p} |\boldsymbol{\varepsilon}^T \mathbf{X}_j|, \frac{1}{n} \max_{j \in \hat{\mathcal{I}}} |\boldsymbol{\varepsilon}^T \hat{\mathbf{W}}_j| \right). \tag{4.9}$$

We have

$$\begin{aligned} &\frac{1}{2n} \left\| \mathbf{X}\left(\hat{\theta} - \theta^*\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{\mathcal{I}}} \hat{\gamma} \right\|_2^2 \\ \leq &\frac{|\hat{\mathcal{I}}^C \cap \mathrm{supp}(\gamma^*)|}{n} \sum_{j \in \hat{\mathcal{I}}^C \cap \mathrm{supp}(\gamma^*)} \left\| \mathbf{W}_j \gamma_j^* \right\|_2^2 + \frac{2}{n} \left\| \left(\mathbf{W}_{\hat{\mathcal{I}}} - \hat{\mathbf{W}}_{\hat{\mathcal{I}}} \right) \gamma_{\hat{\mathcal{I}}}^* \right\|_2^2 + 2\lambda \left(\|\theta^*\|_1 + \|\gamma_{\hat{\mathcal{I}}}^*\|_1 \right). \end{aligned} \tag{4.10}$$

Proof. See Section C.2. □

The deterministic prediction error (4.10) is very general in that (a) it allows for any $\hat{\mathcal{I}}$ from Step 2, and it holds for any value of α in (4.7); (b) it allows for any predicted pure interactions $\hat{\mathbf{W}}$; and (c) it does not specify the distribution of each column of $\mathbf{X}$ and $\hat{\mathbf{W}}$. A careful examination of (4.10) reveals that the success of the final step hinges on these three aspects. Conditional on the success of Step

2, i.e., $\mathcal{I} \subseteq \hat{\mathcal{I}}$, we have $\hat{\mathcal{I}}^C \cap \text{supp}(\gamma^*) \subseteq \mathcal{I}^C \cap \text{supp}(\gamma^*)$. The scale of the first term hinges on the value of α in (4.7). Intuitively, a small value of α leads to a small set $\hat{\mathcal{I}}^C \cap \text{supp}(\gamma^*)$. The second term being small depends on $\hat{\mathbf{W}}_{\hat{\mathcal{I}}}$ being a good approximation of $\mathbf{W}_{\hat{\mathcal{I}}}$. Finally, the scale of λ that satisfies (4.9) determines the rate of the third term in (4.10), and it depends on the distribution of $\mathbf{X}$ as well as $\hat{\mathbf{W}}$. We will see this discussion more specifically in later results. First we consider the setting where $\mathbf{W}$ is available, i.e., when $\hat{\mathbf{W}} = \mathbf{W}$.

Corollary 22. *Suppose **A1** holds with $n^{2\kappa} = o(p)$. Conditional on Theorem 20 succeeding, and suppose $\hat{\mathbf{W}} = \mathbf{W}$, if we take*

$$\lambda = 2\sigma \sqrt{\frac{\max_j \Sigma_{jj} \log p}{cn}}, \tag{4.11}$$

for some constant $c > 0$, then

$$\frac{1}{2n} \left\| \mathbf{X}\left(\hat{\theta} - \theta^*\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{\mathcal{I}}}\hat{\gamma} \right\|_2^2 \leq 2|\hat{\mathcal{I}}^C \cap \text{supp}(\gamma^*)|^2\alpha + 2\lambda\left(\|\theta^*\|_1 + \|\gamma^*_{\hat{\mathcal{I}}}\|_1\right) \tag{4.12}$$

holds with probability greater than $1 - K_1 \exp(-K_2 \log p) - K_3 \exp(-K_4 n^{1/2})$ for some positive constants K_1 through K_4.

Proof. See Section C.3. □

Recall from (4.8) that $|\hat{\mathcal{I}}| = O(n^{2\kappa})$. The assumption $n^{2\kappa} = o(p)$ requires that the size of retained interactions after Step 2 should not be too large in comparison with p.

Although computationally much more expensive, APL enjoys the same prediction error rate as in (4.12) with $\alpha = O(\lambda)$ and λ in (4.11). Corollary 22 implies that the proposed method has a theoretical prediction error rate that is as good as APL if we have access to $\mathbf{W}$, while being much more computationally efficient.

In many settings $\mathbf{W}$ is not available, and an approximation $\hat{\mathbf{W}}$ is needed. The next corollary implies that essentially the same error rate is attained as long as $\hat{\mathbf{W}}$ is a good enough approximation of $\mathbf{W}$.

Corollary 23. *Suppose **A1** holds with $n^{2\kappa} = o(p)$. Conditional on Theorem 20 succeeding, and suppose for some constant $C > 0$, $\|\hat{\mathbf{W}}_j\|_2^2 \leq C\|\mathbf{W}_j\|_2^2$ with probability greater than $1 - h(n)$ with $h(n) \to 0$ as $n \to \infty$. Then if we take*

$$\lambda = 2\sigma\sqrt{\frac{\max_j \Sigma_{jj} \log p}{cn}},$$

for some constant $c > 0$, then

$$\frac{1}{2n}\left\|\mathbf{X}\left(\hat{\theta} - \theta^*\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{\mathcal{I}}}\hat{\gamma}\right\|_2^2 \leq 2|\hat{\mathcal{I}}^C \cap \mathrm{supp}(\gamma^*)|^2\alpha + \frac{2}{n}\left\|\left(\mathbf{W}_{\hat{\mathcal{I}}} - \hat{\mathbf{W}}_{\hat{\mathcal{I}}}\right)\gamma^*_{\hat{\mathcal{I}}}\right\|_2^2 + 2\lambda\left(\|\theta^*\|_1 + \|\gamma^*_{\hat{\mathcal{I}}}\|_1\right)$$

holds with probability greater than $1 - h(n) - K_1\exp(-K_2\log p) - K_3\exp(-K_4 n^{1/2})$ for some positive constants K_1 through K_4.

Proof. See Section C.4. □

For the proposed method to have the same prediction error rate (in p and n) as APL, we require $n^{-1}\|(\mathbf{W}_{\hat{\mathcal{I}}} - \hat{\mathbf{W}}_{\hat{\mathcal{I}}})\gamma^*_{\hat{\mathcal{I}}}\|_2^2$ to be of order $(n^{-1}\log p)^{1/2}$. For example, if $\hat{\mathbf{W}}_{\hat{\mathcal{I}}}$ attains $\|\hat{\mathbf{W}}_{\hat{\mathcal{I}}} - \mathbf{W}_{\hat{\mathcal{I}}}\|_{op} = O\{(n^{-1}\log p)^{1/2}\}$, where $\|\cdot\|_{op}$ is the matrix operator norm, then the desired prediction error rate as that of APL is achieved.

Our final corollary gives a specific example of (4.10) where one uses the lasso to get $\hat{\mathbf{W}}_{\hat{\mathcal{I}}}$, where each column $\hat{\mathbf{W}}_j = \mathbf{Z}_j - \mathbf{X}\hat{\phi}_j$ for $j \in \hat{\mathcal{I}}$, and $\hat{\phi}_j$ is a solution to the following lasso problem (4.5).

Corollary 24. *Suppose **A1** holds with $n^{2\kappa} = o(p)$. Conditional on Theorem 20 succeeding, for $\mathcal{I}$ defined in* (4.7)*, if we take*

$$\lambda \geq 2\sigma\sqrt{\frac{\max_j \Sigma_{jj}\log p}{c_1 n}} \quad \textit{and} \quad \nu \geq \sqrt{\frac{2\log p + \log|\hat{\mathcal{I}}|}{c_2 n^{2/3}}}$$

for some constants $c_1, c_2 > 0$, *then*

$$\frac{1}{2n}\left\|\mathbf{X}\left(\hat{\theta}-\theta^*\right)+\mathbf{W}\gamma^*-\hat{\mathbf{W}}_{\hat{\mathcal{I}}}\hat{\gamma}\right\|_2^2 \leq 2|\hat{\mathcal{I}}^C\cap \mathrm{supp}(\gamma^*)|^2\alpha + 2\nu\sum_{j\in\hat{\mathcal{I}}}\|\Sigma^{-1}\Phi_j\|_1\|\gamma^*_{\hat{\mathcal{I}}}\|_2^2 + 2\lambda\left(\|\theta^*\|_1+\|\gamma^*_{\hat{\mathcal{I}}}\|_1\right)$$

holds with probability greater than $1-h(n)-K_1\exp(-K_2\log p)-K_3\exp(-K_4 n^{1/2})$ *for some positive constants* K_1 *through* K_4.

Proof. See Section C.5. □

We observe the extra price to be paid for fitting a lasso to get $\hat{\mathbf{W}}$. In particular, the rate in $O\{(n^{-2/3}\log p)^{1/2}\|\gamma^*\|_2^2\}$ is slower than that of the APL. The main difficulty leading to this rate is the "empirical process" part (Bühlmann & Van De Geer 2011) in (4.5):

$$(\hat{\phi}^{(j)}-\phi^{(j)})^T\mathbf{X}^T\mathbf{Z}_j \leq \max_{k\in\hat{\mathcal{I}}}|\mathbf{X}_k^T\mathbf{Z}_j|\cdot\|\hat{\phi}^{(j)}-\phi^{(j)}\|_1.$$

In particular, the distribution of each column of $\mathbf{Z}$ has a heavier tail than the main effects, and the sample quantities based on $\mathbf{W}$ concentrate more slowly to their corresponding population quantities.

4.4.1 On Assumption A3

Although the proposed framework does not depend on a specific regression method in Step 1 for fitting the main effects, we take the lasso as an example. Recall that we are fitting a lasso with a misspecified model, i.e., we treat $W^T\gamma^*+\varepsilon$ in (4.3) as the noise term and solve the following problem:

$$\check{\theta}\in\arg\min_{\theta}\left(\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\theta\|_2^2+\lambda\|\theta\|_1\right).$$

The following theorem gives a prediction error rate for the main effects only lasso that is carried out in Step 1.

Theorem 25. *Suppose that Assumption **A1** holds. For any $t^2 > 0$, take*

$$\lambda_0 = K_1\sigma\sqrt{\frac{t^2 + \log p}{n}} + K_2\|\gamma^*\|_1\sqrt{\frac{t^2 + \log p}{n^{2/3}}}$$

for some constants $K_1, K_2 > 0$, the following prediction error rate for Step 1 holds with probability greater than $1 - C_1\exp(1 - t^2) - C_3\exp(\log p - C_2 n)$ for some constants $C_1, C_2, C_3 > 0$:

$$\frac{1}{2n}\|\mathbf{X}\check{\theta} - \mathbf{X}\theta^*\|_2^2 \leq 2\lambda\|\theta^*\|_1.$$

Proof. See Section C.6. □

We see that the presence of interactions leads to rates that are less good than if no interactions were present. This is the price paid for fitting a misspecified model in Step 1. The slower rate is due to dealing with the empirical process that involves the heavier tailed interactions.

Yet it can be checked that if $\|\theta^*\|_1 = O(n^{\xi_2})$ for some $\xi_2 > 0$, then for $\tau \leq \frac{1}{6} - \frac{\xi}{4} - \frac{\xi_1+\xi_2}{2}$ we have **A3** holds. Furthermore, under stronger conditions (e.g., compatibility conditions on θ^*), a faster prediction error rate in Step 1 could be derived.

4.4.2 Comparison of Theorem 20 with other methods

As with other methods in interaction screening, we note that the result in Theorem 20 is less favorable than that of the sure independence screening (Fan & Lv 2008) when directly applied to the main effects, which can handle problems when $\xi + 2\kappa < 1$. This reveals the intrinsic challenge when dealing with interactions, which have heavier tails than main effects. When we further assume that

X has a bounded distribution, Theorem 20 can be much improved. Note that this is still weaker than that in Fan & Lv (2008), as it depends on the misspecified lasso fit in the first step—an expected caveat in a two-stage method.

It is of interest to compare the results in Theorem 20 with other two-stage methods. The forward-selection based two-stage methods iFORT and iFORM proposed in Hao & Zhang (2014) require hierarchical structure within interactions and provides screening property guarantees. Our method, while not requiring hierarchy, achieves the same screening property with similar assumptions.

Fan et al. (2016) considered screening based on "active interaction variables" and constructed interactions within the selected active interaction variables. The method, called Interaction Pursuit (IP), is particularly effective when most interactions are based on a small number of "active" main effects. Again, Step 2 achieves the same screening property as IP under similar assumptions, while we do not put any assumptions on the underlying structures among interactions.

4.5 Numerical studies

4.5.1 Simulation studies: binary features

We consider a simulation scenario with binary features in which interactions can be well approximated by main effects. We generate p binary features as follows: For a (perfect) binary tree of depth $d = 5$, each leaf node is an independent Bernoulli(0.1) random variable; the value of each non-leaf node is the

maximum of the node values in its sub-tree, i.e., each non-leaf node represents an event that is the union of all the events represented by its children nodes. The total number of nodes in the tree is $p = 2^{d+1} - 1$, and we consider these node values as main effects. This construction ensures that for any pair of main effects, they are either independent or else one is an ancestor of the other. The interaction between two binary features is simply the intersection of the two main effect events, so in this second case their interaction is simply the main effect corresponding to the descendent node. Figure 4.1 shows the binary tree (of depth 5), where each node representing a main effect, and the node value is the success probability of the corresponding Bernoulli random variable.

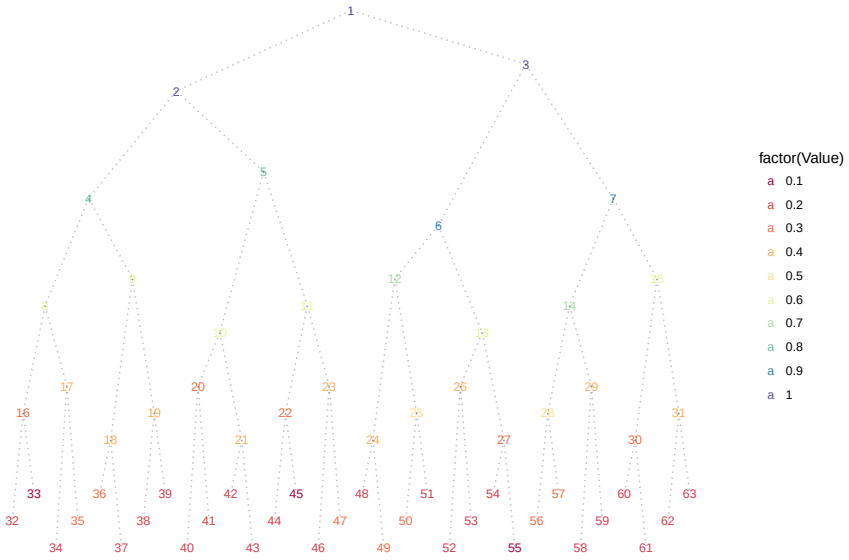

Figure 4.1: An example of the perfect binary tree, representing main effects. Node value represents the success probability (rounded to 1 decimal place) of the corresponding Bernoulli random variable.

We can control the degree to which the interaction signal can be explained by main effects by choosing the proportion of nonzero elements of γ^* correspond-

ing to interactions between main effects that are ancestors/descendants of each other versus not. We consider three scenarios: (a) almost all interactions can be explained by main effects; (b) approximately half of the interactions can be explained by main effects; and (c) a very limited amount of interactions can be explained by main effects. These three scenario correspond to three cases where the *main-effect-interaction-ratio*,

$$\text{mir} = \frac{\|\mathbf{X}\theta^*\|_2^2}{\|\mathbf{W}\gamma^*\|_2^2},$$

is large, medium, and small. For each value of mir, we generate the response $\mathbf{y}$ using (4.3) with the zero-mean additive noise ε generated according to the signal-to-noise ratio $\frac{\|\mathbf{X}\theta^*\|^2+\|\mathbf{W}\gamma^*\|^2}{n\sigma^2} \in \{0.5, 1, 2, 3, 4\}$. We generate $n = 100$ samples in each simulation setting, and in Figure 4.2 we report the prediction error of various methods (averaged over 100 repetitions). In particular, we compare the performance of the following methods:

- The all pairs lasso (APL) with tuning parameter selected by cross-validation.
- The main effects Lasso (MEL) with tuning parameter selected by cross-validation.
- Oracle: Least squares estimate with an oracle knowledge of true support.
- `sprinter_w`, as in Algorithm 4, with lasso using main effects in Step 1. We use cross-validation in Step 1 before going to subsequent steps, and an additional cross-validation is used in Step 2 and 3 together to select the final model. In Step 3, we use $\hat{\mathbf{W}}$ as computed in (4.5).
- `sprinter_z`, as in Algorithm 4, with lasso using main effects in Step 1. We use cross-validation in Step 1 before going to subsequent steps, and

an additional cross-validation is used in Step 2 and 3 together to select the final model. In Step 3, we use $\hat{\mathbf{W}} = \mathbf{Z}$.

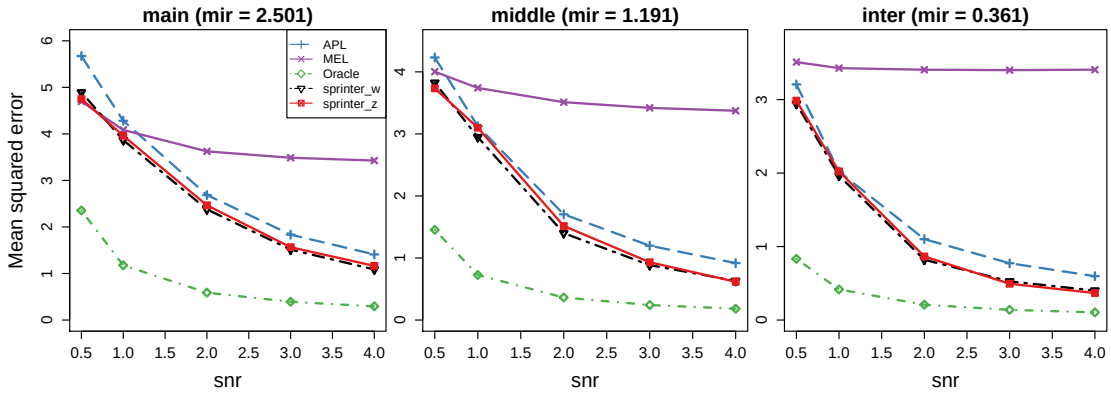

Figure 4.2: Prediction mean-squared error of different methods (averaged over 100 repetitions, binary settings).

As mir gets small, the performance of MEL worsens relative to other methods that model interactions. The performance of both versions of sprinter are favorable in comparison with APL and MEL. Moreover, we find very little difference in prediction error by using different $\hat{\mathbf{W}}$ in Step 3.

4.5.2 Simulation studies: Gaussian features

We generate $n = 100$ samples from model (4.1), where X is a p-dimensional random vector following a multivariate Gaussian distribution with mean $\mathbf{0}$, $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \leq j, k \leq p$, and $p = 400$. Recall that the principle in Section 4.2 is different with the hierarchical structure among interactions, and the proposed method does not assume hierarchy; Actually sprinter does not assume any structure among interactions. And we study the performance of

sprinter in different interaction structures. Denote $\mathcal{T}_1$ as the indices of main effects, $\mathcal{T}_2$ as the indices of squared terms and $\mathcal{T}_3$ as indices of interaction terms, and consider the following structures for the interaction effects Z:

1. Mixed: $\mathcal{T}_1 = \{1, 2, \ldots, 6\}, \mathcal{T}_2 = \{1, 5, 15\},$
   $\mathcal{T}_3 = \{(1, 5), (4, 18), (10, 11), (9, 17), (1, 13), (4, 17)\}$.
2. Hierarchical, i.e., $\beta_{jk} \neq 0 \implies \beta_j \neq 0$ or $\beta_k \neq 0$: $\mathcal{T}_1 = \{1, 2, \ldots, 6\}$, $\mathcal{T}_2 = \{1, 2, 3\}$ and $\mathcal{T}_3 = \{(1, 3), (2, 4), (3, 4), (1, 8), (2, 8), (5, 10)\}$.
3. Anti-hierarchical, i.e., $\beta_{jk} \neq 0 \implies \beta_j = 0, \beta_k = 0$: $\mathcal{T}_1 = \{1, 2, \ldots, 6\}$, $\mathcal{T}_2 = \{11, 12, 13\}$ and
   $\mathcal{T}_3 = \{(11, 13), (12, 14), (13, 14), (11, 18), (12, 18), (15, 20)\}$.
4. Interaction only: $\mathcal{T}_1 = \mathcal{T}_2 = \emptyset$ and $\mathcal{T}_3 = \{(1, 3), (2, 4), (3, 4), (1, 8), (2, 8), (5, 10)\}$.
5. Main effects only: $\mathcal{T}_1 = \{1, 2, \ldots, 6\}$, $\mathcal{T}_2 = \emptyset$ and $\mathcal{T}_3 = \emptyset$.
6. Squared effects only: $\mathcal{T}_1 = \emptyset$, $\mathcal{T}_2 = \{1, 2, \ldots, 6\}$ and $\mathcal{T}_3 = \emptyset$.

Note that interaction hierarchy structure only exists in the hierarchical model and the main effects only model. The signal strength is then set as $\beta_j^* = 2$ for $j \in \mathcal{T}_1$, $\gamma_j^* = 3$ for $j \in \mathcal{T}_2$ and $j \in \mathcal{T}_3$. Finally, the zero-mean additive noise ε in (4.1) is generated according to the signal-to-noise ratio $\sqrt{\frac{\|\mathbf{X}\beta^*\|^2 + \|\mathbf{Z}\gamma^*\|^2}{n\sigma^2}} \in \{0.5, 1, 2, 3, 4, 5\}$.

Recall that in the Gaussian case, $\mathbf{W} = \mathbf{Z}$. So `sprinter_z` is essentially the same as `sprinter_w`, which we will simply call `sprinter`. In addition to the methods considered in previous study, we also include the performance of the following methods:

- RAMP (Hao et al. 2018), which iteratively adds variables into a path of solutions under marginality (hierarchy) principle. They also consider the

two-stage lasso, but state that RAMP performs better than the two-stage lasso (Hao & Zhang 2014).

- Interaction Pursuit (IP) by Fan et al. (2016).
- SIS + Lasso: We use sure independence screening (Fan & Lv 2008) on all main effects and interactions, and fit the lasso on the selected candidate features.
- Oracle: Least squares estimate with an oracle knowledge of true support.

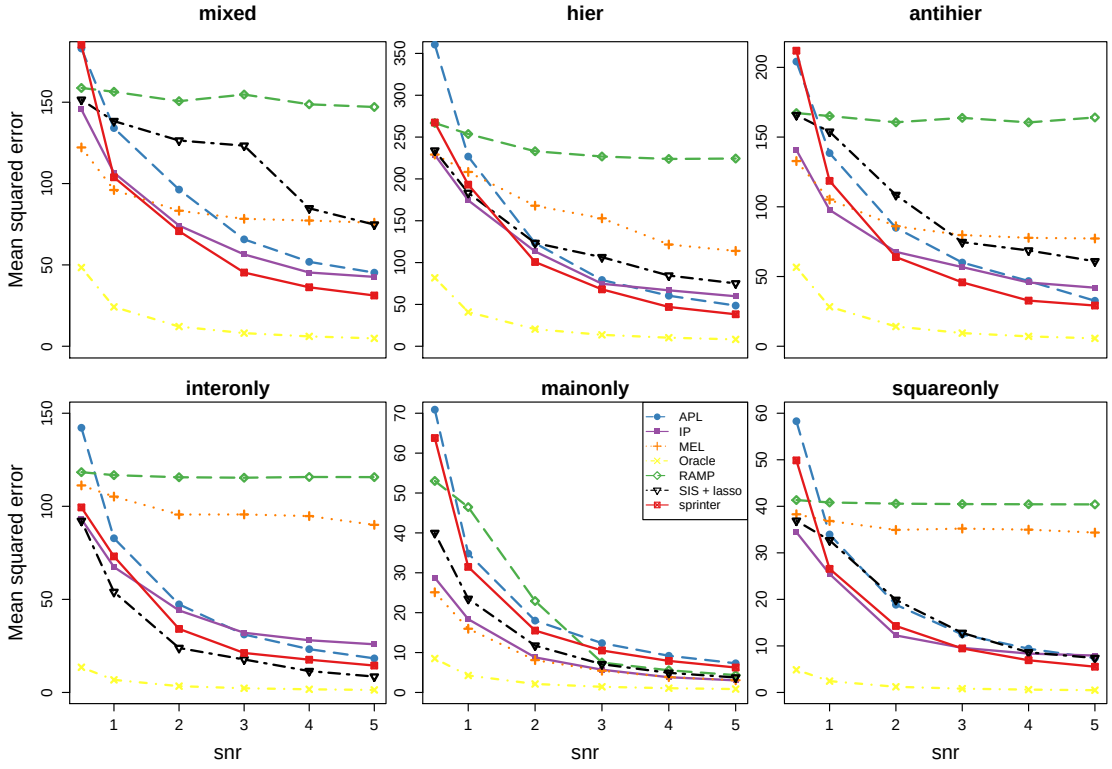

Figure 4.3: Prediction mean-squared error of different methods (averaged over 100 repetitions, Gaussian settings)

We measure the statistical performance of each methods in prediction error, which is averaged over 100 repetitions and is reported in Figure 4.3. Observe

that sprinter almost works uniformly better than other methods in all settings except the main effects only model. This is because sprinter include both main effects and the squared effects in Step 1, which has already been a misspecified model if the response depends on main effects only. Actually, sprinter works much better in this setting if it uses only main effects in Step 1.

4.5.3 Simulation studies: computation time

In this section, we show that sprinter is much more computationally efficient than APL, while having similar (if not better) statistical performance. To this end, we consider varying $p \in \{100, 200, 400, 1000, 2000\}$ in the mixed model in Section 4.5.2 with signal-to-noise ratio equal to 3 and $n = 100$. The following plots show both the computation time (in seconds) and the prediction mean squared error (averaged over 10 repetitions).

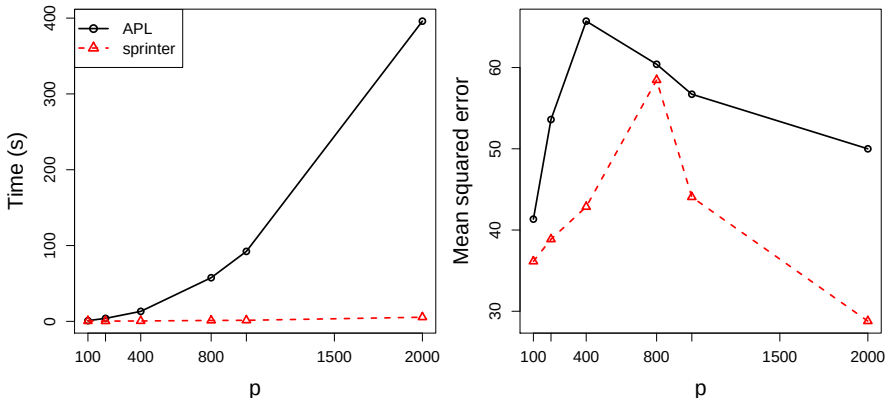

Figure 4.4: Computation time and prediction mean-squared error for different p in the mixed model.

As expected, APL is computationally much more expensive than the proposed method. In particular, for $p = 2000$, the proposed method is about 100

times faster than APL. In addition, while not shown, sprinter can solve a problem with 140000 main effects (about 10 billion interactions, which is infeasible for APL) with 5-fold cross-validation under 7 hours on a single CPU.

In addition to enjoying obvious computational benefits, the right panel of Figure 4.4 shows that the proposed method does not lose statistical performance in terms of prediction error. Actually, sprinter attains even smaller prediction error than APL.

4.5.4 Data example: Riboflavin

Finally, we consider applying sprinter to the Riboflavin data set (Bühlmann et al. 2014), which is also considered in Thanei et al. (2016). The data set contains $p = 4088$ gene-expression features and $n = 71$ observations, which are randomly split into set A of size 30 and set B of size 31. We first use set A as the training set and set B as the testing set, and then we reverse the roles of A and B. To measure the statistical performance, we report the normalized out-of-sample prediction error (Thanei et al. 2016):

$$r^2 = \frac{\|\mathbf{y}_{\text{test}} - \hat{\mathbf{y}}_{\text{test}}\|^2}{\|\mathbf{y}_{\text{test}}\|^2}.$$

The following table shows both the computing time and r^2 for sprinter, the xyz algorithm (Thanei et al. 2016), and APL.

	xyz	APL	sprinter
Time (s)	13.5285	773.5015	11.4360
r^2	1.59274	0.01267	0.00935

While sprinter is about 70 times faster than APL, it achieves an even higher r^2. By contrast, xyz is about as efficient as sprinter, but suffers from poor prediction performance.

CHAPTER 5

CONCLUSION

This dissertation proposes three methods, each of which studies a component of a general framework of the high-dimensional structured regression problem. A recurring theme of the proposed methods is that they cast a certain structured statistical problem as a convex optimization problem.

In Chapter 2, we propose a new flexible method for learning local dependence in the setting where the elements of a random vector have a known ordering. The model amounts to sparse estimation of the inverse of the Cholesky factor of the covariance matrix with variable bandwidth. Our method is based on a convex formulation that allows it to simultaneously yield a flexible adaptively-banded sparsity pattern, enjoy efficient computational algorithms, and be studied theoretically. To our knowledge, no previous method has all these properties. We show how the matrix estimation problem can be decomposed into independent row estimation problems, each of which can be solved via an ADMM algorithm having efficient updates. We prove that our method recovers the signed support of the true Cholesky factor and attains estimation consistency rates in several matrix norms under assumptions as mild as those in linear regression problems. Simulation studies show that our method compares favorably to two pre-existing estimators in the ordered setting, both in terms of support recovery and in terms of estimation accuracy. Through a genetic data example, we illustrate how our method may be applied to model the local dependence of genetic variations in genes along a chromosome. Finally, we illustrate that our method has favorable performance in a sound recording classification problem.

Chapter 3 considers the problem of estimating the error variance in a high-dimensional linear model. Compared with the vast literature in estimating the regression coefficients, the problem of estimating error variance in high-dimensional linear model has been under-develeoped, especially relative to its importance in high-dimensional statistical inference. We show that the natural parametization of the multiparameter exponential family of a Gaussian with unknown mean and variance leads to a new approach to this problem with attractive properties. Our new estimators, natural lasso and organic lasso, admit finite sample bounds that do not make any assumptions on the design matrix. From a practical standpoint, our estimators are easy to compute and have empirical performance surpassing existing approaches. Another contribution of Chapter 3 is to provide a more complete view of high-dimensional error variance estimators under no conditions on the design matrix. In particular, we show that two popular pre-existing estimators of error variance share the same error rate with the proposed methods under no conditions on the design matrix. To our best knowledge, these results are not found in previous literature. An interesting theoretical extension of Chapter 3 is to study if the proposed methods attain a faster rate under stronger assumptions. And another important future step would be the application of the proposed methods in certain high-dimensional inferential tasks.

In Chapter 4, we propose a new principle (RISP) in large-scale interaction modeling, which says that one should prefer main effects over interactions given similar prediction performance. The proposed method, sprinter, is a multiple-stage method that honors RISP: in the first stage it tries to capture as much of the variability in the response as possible without resorting to interactions; in the second stage it includes only interactions that capture signal

that cannot be captured by main effects. In this sense, sprinter is a reluctant interaction selection procedure. The reluctance of sprinter allows for interaction modeling on unprecedented problem sizes without compromising practical statistical performance. Extensive numerical studies are carried out to show that sprinter performs favorably, both in prediction error and (espcially) in computational efficiency. In addition, sprinter attains theoretical properties that compare favorably with alternative methods that are also computationally efficient.

Actually, the principle in Chapter 4 is not limited to interaction modeling. It can be applied in a more general setting where one would give preference to one set of variables A (in analogy to the main effects) over the other set B (in analogy to the interactions). For example, A could be the set of cheap and easy-to-get features while features in set B are much more expensive. Or B could be the set of (high-dimensional) confounding variables of A. In both examples, it is an important yet challenging problem to understand how to capture as much of the variability in the response as possible without resorting to features in B. It is also very interesting to understand the pure effect of features in B in the presence of A. This could be a natural and important extension of Chapter 4.

APPENDIX A

APPENDIX OF CHAPTER 2

A.1 Decoupling property

Let $S = \frac{1}{n}\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p\times p}$ be the sample covariance matrix. Then the estimator (2.5) is the solution to the following minimization problem:

$$\min_{\substack{L:L_{rr}>0\\ L_{rk}=0 \text{ for } r<k}} \left\{ -2\sum_{r=1}^{p}\log L_{rr} + \operatorname{tr}(S\,L^T L) + \lambda \sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 L_{rm}^2} \right\}.$$

First note that under the lower-triangular constraint

$$\operatorname{tr}\left(S\,L^T L\right) = \frac{1}{n}\sum_{r=1}^{p}\operatorname{tr}\left(\mathbf{X}L_{\cdot r}^T L_{r\cdot}\mathbf{X}^T\right) = \frac{1}{n}\sum_{r=1}^{p}\left\|\mathbf{X}L_{\cdot r}^T\right\|_2^2 = \frac{1}{n}\sum_{r=1}^{p}\left\|\mathbf{X}_{1:r}L_{1:r,r}^T\right\|_2^2,$$

where $\mathbf{X}_{1:r}$ is a matrix of the first r columns of $\mathbf{X}$. Thus

$$\begin{aligned}
&-2\sum_{r=1}^{p}\log L_{rr} + \operatorname{tr}(S\,L^T L) + \lambda\sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 L_{rm}^2} \\
&= -2\log L_{11} + \frac{1}{n}\left\|\mathbf{X}_1 L_{11}\right\|_2^2 + \sum_{r=2}^{p}\left(-2\log L_{rr} + \frac{1}{n}\left\|\mathbf{X}_{1:r}L_{1:r,r}^T\right\|_2^2 + \lambda\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 L_{rm}^2}\right).
\end{aligned}$$

Therefore the original problem can be decoupled into p separate problems. In particular, a solution $\hat{L}$ can be written in a row-wise form with

$$\hat{L}_{11} = \operatorname*{arg\,min}_{L_{11}>0}\left\{-2\log L_{11} + \frac{1}{n}\left\|\mathbf{X}_1 L_{11}\right\|_2^2\right\} = \frac{1}{\sqrt{S_{11}}},$$

and for $r = 2, \ldots, p$,

$$\hat{L}_{1:r,r}^T = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^r:\beta_r>0}\left\{-2\log\beta_r + \frac{1}{n}\left\|\mathbf{X}_{1:r}\beta\right\|_2^2 + \lambda\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2\beta_m^2}\right\}.$$

A.2 A closed-form solution to (2.9)

The objective function in (2.9) is a smooth function. Taking the derivative with respect to β and setting to zero gives the following system of equations:

$$-2\frac{1}{\beta_r}\mathbf{e}_r + \frac{2}{n}\mathbf{X}_{1:r}^T\mathbf{X}_{1:r}\beta + u^{(t-1)} + \rho\left(\beta - \gamma^{(t-1)}\right) = \mathbf{0}.$$

Letting $S^{(r)} = \frac{1}{n}\mathbf{X}_{1:r}^T\mathbf{X}_{1:r}$, then the equations above can be further decomposed into

$$\begin{aligned}
&-\frac{2}{\beta_r} + \left(2S_{rr}^{(r)} + \rho\right)\beta_r + 2S_{r,-r}^{(r)}\beta_{-r} + u_r^{(t-1)} - \rho\gamma_r^{(t-1)} = 0,\\
&\left(2S_{-r,-r}^{(r)} + \rho I\right)\beta_{-r} + 2S_{-r,r}^{(r)}\beta_r + u_{-r}^{(t-1)} - \rho\gamma_{-r}^{(t-1)} = \mathbf{0}.
\end{aligned}$$

Solving for β_{-r} in the second system of equations gives

$$\beta_{-r} = -\left(2S_{-r,-r}^{(r)} + \rho I\right)^{-1}\left(2S_{-r,r}^{(r)}\beta_r + u_{-r}^{(t-1)} - \rho\gamma_{-r}^{(t-1)}\right),$$

which is then plugged back in the first equation to give

$$2\frac{1}{\beta_r} + A\beta_r + B = 0,$$

where

$$\begin{aligned}
A &= 4S_{r,-r}^{(r)}\left(2S_{-r,-r}^{(r)} + \rho I\right)^{-1}S_{-r,r}^{(r)} - 2S_{r,r}^{(r)} - \rho,\\
B &= 2S_{r,-r}^{(r)}\left(2S_{-r,-r}^{(r)} + \rho I\right)^{-1}\left(u_{-r}^{(t-1)} - \rho\gamma_{-r}^{(t-1)}\right) - u_r^{(t-1)} + \rho\gamma_r^{(t-1)}.
\end{aligned}$$

Solving for β_r gives the closed-form update.

A.3 Dual problem of (2.10)

Lemma 26. *A dual problem of* (2.10) *is*

$$\min_{a^{(\ell)}\in\mathbb{R}^r}\left\{\left\|y^{(t)} - \frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}W^{(\ell)} * a^{(\ell)}\right\|_2^2 \quad \text{s.t.} \quad \left\|\left(a^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 \le 1, \quad \left(a^{(\ell)}\right)_{g_{r,\ell}^c} = 0\right\}, \tag{A.1}$$

where $y^{(t)} = \beta^{(t)} + \frac{1}{\rho}u^{(t-1)}$. *Also, given a solution* $\hat{a}^{(1)}, \ldots, \hat{a}^{(r-1)}$, *the solution to* (2.10) *can be written as*

$$\gamma^{(t)} = y^{(t)} - \frac{\lambda}{\rho}\sum_{\ell=1}^{r-1} W^{(\ell)} * \hat{a}^{(\ell)}. \tag{A.2}$$

Proof. Note that

$$\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \gamma_m^2} = \left\| \left(W^{(\ell)} * \gamma \right)_{g_{r,\ell}} \right\|_2$$
$$= \max \left\{ \left\langle W^{(\ell)} * a^{(\ell)}, \gamma \right\rangle, \quad \text{s.t.} \quad \left\| \left(a^{(\ell)} \right)_{g_{r,\ell}} \right\|_2 \leq 1, \quad \left(a^{(\ell)} \right)_{g_{r,\ell}^c} = 0 \right\}.$$

Thus, the minimization problem in (2.10) becomes

$$\min_{\gamma} \left\{ \frac{1}{2} \left\| \gamma - y^{(t)} \right\|_2^2 + \frac{\lambda}{\rho} \sum_{\ell=1}^{r-1} \left\| \left(W^{(\ell)} * \gamma \right)_{g_{r,\ell}} \right\|_2 \right\}$$
$$= \min_{\gamma} \left\{ \max_{a^{(\ell)}} \left\{ \frac{1}{2} \left\| \gamma - y^{(t)} \right\|_2^2 + \frac{\lambda}{\rho} \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * a^{(\ell)}, \gamma \right\rangle, \left\| \left(a^{(\ell)} \right)_{g_{r,\ell}} \right\|_2 \leq 1, \left(a^{(\ell)} \right)_{g_{r,\ell}^c} = 0 \right\} \right\}$$
$$= \max_{a^{(\ell)}} \left\{ \min_{\gamma} \left\{ \frac{1}{2} \left\| \gamma - y^{(t)} \right\|_2^2 + \frac{\lambda}{\rho} \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * a^{(\ell)}, \gamma \right\rangle, \left\| \left(a^{(\ell)} \right)_{g_{r,\ell}} \right\|_2 \leq 1, \left(a^{(\ell)} \right)_{g_{r,\ell}^c} = 0 \right\} \right\},$$

where $y^{(t)} = \beta^{(t)} + \frac{1}{\rho}u^{(t-1)}$. We solve the inner minimization problem by setting the derivative to zero,

$$\gamma - y^{(t)} + \frac{\lambda}{\rho} \sum_{\ell=1}^{r-1} W^{(\ell)} * a^{(\ell)} = 0,$$

which gives the primal-dual relation,

$$\gamma = -\frac{\lambda}{\rho} \sum_{\ell=1}^{r-1} W^{(\ell)} * a^{(\ell)} + y^{(t)}.$$

Using this gives

$$
\begin{aligned}
&\min_{\gamma}\left\{\frac{1}{2}\left\|\gamma-y^{(t)}\right\|_2^2+\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}\left\|\left(W^{(\ell)}*\gamma\right)_{g_{r,\ell}}\right\|_2\right\}\\
&=\max_{a^{(\ell)}}\left\{\frac{1}{2}\left\|-\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}W^{(\ell)}*a^{(\ell)}\right\|_2^2+\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}\left\langle W^{(\ell)}*a^{(\ell)},-\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}W^{(\ell)}*a^{(\ell)}+y^{(t)}\right\rangle\right.\\
&\qquad\left.\text{s.t.}\quad\left\|\left(a^{(\ell)}\right)_{g_{r,\ell}}\right\|_2\le 1,\quad\left(a^{(\ell)}\right)_{g_{r,\ell}^c}=0\right\}\\
&=\min_{a^{(\ell)}}\left\{\left\|y^{(t)}-\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}W^{(\ell)}*a^{(\ell)}\right\|_2^2\quad\text{s.t.}\quad\left\|\left(a^{(\ell)}\right)_{g_{r,\ell}}\right\|_2\le 1,\quad\left(a^{(\ell)}\right)_{g_{r,\ell}^c}=0\right\}.
\end{aligned}
$$

□

Algorithm 5: BCD on the dual problem (A.1)

1: Let $y^{(t)}=\beta^{(t)}+\frac{1}{\rho}u^{(t-1)}$

2: Initialize $\hat{a}^{(\ell)}\leftarrow 0$ for all $\ell=1,\cdots,r-1$

3: **for** $\ell=1,\cdots,r-1$ **do**

4: $\quad\hat{z}^{(\ell)}\leftarrow y^{(t)}-\frac{\lambda}{\rho}\sum_{k=1}^{r-1}W^{(k)}*\hat{a}^{(k)}$ Find a root $\hat{\nu}_\ell$ that satisfies

$$h_\ell(\nu):=\sum_{m=1}^{\ell}\frac{w_{\ell m}^2}{\left(w_{\ell m}^2+\nu\right)^2}\left(\hat{z}_m^{(\ell)}\right)^2=\frac{\lambda^2}{\rho^2}\tag{A.3}$$

5: $\quad$**for** $m=1,\cdots,\ell$ **do**

6: $\quad\quad\hat{a}_m^{(\ell)}\leftarrow\frac{w_{\ell m}}{\frac{\lambda}{\rho}\left(w_{\ell m}^2+[\hat{\nu}_\ell]_+\right)}\hat{z}_m^{(\ell)}$

7: **return** $\left\{\hat{a}^{(\ell)}\right\}$ as a solution to (A.1)

8: **return** $\gamma^{(t)}=y^{(t)}-\frac{\lambda}{\rho}\sum_{\ell=1}^{r-1}W^{(\ell)}*\hat{a}^{(\ell)}$ as a solution to (2.10)

A.4 Elliptical projection

We adapt the same procedure as in Appendix B of Bien et al. (2016) to update one $a^{(\ell)}$ in Algorithm (5). By (2.10) we need to solve a problem of the form

$$\min_{a \in \mathbb{R}^\ell} \left\| \hat{z}^{(\ell)} - \tau D a \right\|_2^2 \quad \text{s.t.} \quad \|a\|_2 \leq 1,$$

where $\tau = \frac{\lambda}{\rho}$ and $D = \operatorname{diag}(w_{\ell m})_{m \leq \ell} \in \mathbb{R}^{\ell \times \ell}$. If $\left\| D^{-1} \hat{z}^{(\ell)} \right\|_2 \leq \tau$, then clearly $\hat{a} = \frac{1}{\tau} D^{-1} \hat{z}^{(\ell)}$. Otherwise, we use the Lagrangian multiplier method to solve the constrained minimization problem above. Specifically, we find a stationary point of

$$\mathcal{L}(a, \nu) = \left\| \hat{z}^{(\ell)} - \tau D a \right\|_2^2 + \nu \tau^2 \left(\|a\|_2^2 - 1 \right).$$

Taking the derivative with respect to a and set it equal to zero, we have

$$\hat{a}_m = \frac{w_{\ell m}}{\tau (w_{\ell m}^2 + \hat{\nu})} \hat{z}_m^{(\ell)},$$

for each $m \leq \ell$, and $\hat{\nu}$ is such that $\|\hat{a}\|_2 = 1$, which means it satisfies (A.3). By observing that $h_\ell(\nu)$ is a decreasing function of ν and $w_{\ell\ell} = \max_{m \leq \ell} w_{\ell m}$, following Appendix B of Bien et al. (2016), we obtain lower and upper bounds for $\hat{\nu}$:

$$\left[\frac{1}{\tau} \left\| D \hat{z}^{(\ell)} \right\|_2 - w_{\ell\ell}^2 \right]_+ \leq \hat{\nu} \leq \frac{1}{\tau} \left\| D \hat{z}^{(\ell)} \right\|_2,$$

which can be used as an initial interval for finding $\hat{\nu}$ using Newton's method. In practice, we usually find $\hat{\nu}$ from the equation $\frac{1}{h(\nu)} = \tau^{-2}$ for better numerical stability.

We end this section with a characterization of the solution to (2.10), which says that the solution can be written as $\gamma^{(t)} = y^{(t)} * \hat{t}$, where $\hat{t}$ is some data-dependent vector in $\mathbb{R}^r$.

Theorem 27. *A solution to* (2.10) *can be written as* $\gamma^{(t)} = y^{(t)} * \hat{g}$*, where the data-dependent vector* $\hat{g} \in \mathbb{R}^r$ *is given by*

$$\hat{g}_m = \prod_{\ell=m}^{r-1} \frac{[\hat{v}_\ell]_+}{w_{\ell m}^2 + [\hat{v}_\ell]_+}$$

and $\hat{g}_r = 1$*, where* $\hat{v}_\ell$ *satisfies* $\tau^2 = \sum_{m=1}^{\ell} \frac{w_{\ell m}^2}{\left(w_{\ell m}^2 + \nu\right)^2} \left(\hat{z}_m^{(\ell)}\right)^2$.

Proof. By Jenatton et al. (2011), we can get a solution to (2.10) in a single pass as described in Algorithm 5. If we start from $\hat{z}^{(1)} = y^{(t)}$, then for $\ell = 1, \cdots, r-1$ and each $m \le \ell$,

$$\hat{z}_m^{(\ell+1)} = \hat{z}_m^{(\ell)} - \tau w_{\ell m} \hat{a}_m^{(\ell)} = \frac{[\hat{v}_\ell]_+}{w_{\ell m}^2 + [\hat{v}_\ell]_+} \hat{z}_m^{(\ell)}.$$

By (A.2), $\gamma^{(t)} = \hat{z}^{(r-1)}$, and the result follows. □

A key observation from this characterization is that a banded sparsity pattern is induced in solving (2.10), which in turn implies the same property of the output of Algorithm 1.

Corollary 28. *A solution* $\gamma^{(t)}$ *to* (2.10) *has banded sparsity, i.e.,* $\left(\gamma^{(t)}\right)_{1:\hat{J}} = 0$ *for* $\hat{J} = \max\{\ell : \hat{v}_\ell \le 0\}$.

A.5 Uniqueness of the sparse row estimator

Lemma 29. *(Optimality condition) For any* $\lambda > 0$ *and a n-by-p sample matrix* $\mathbf{X}$*,* $\hat{\beta}$ *is a solution to the problem*

$$\min_{\beta \in \mathbb{R}^r} \left\{ -2 \log \beta_r + \frac{1}{n} \|\mathbf{X}_{1:r}\beta\|_2^2 + \lambda \sum_{\ell=1}^{r-1} \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \beta_m^2} \right\}$$

if and only if there exist $\hat{a}^{(\ell)} \in \mathbb{R}^r$ *for* $\ell = 1, \ldots, r-1$ *such that*

$$-\frac{2}{\hat{\beta}_r}\mathbf{e}_r + \frac{2}{n}\mathbf{X}_{1:r}^T\mathbf{X}_{1:r}\hat{\beta} + \lambda \sum_{\ell=1}^{r-1} W^{(\ell)} * \hat{a}^{(\ell)} = 0 \tag{A.4}$$

with $\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}^c} = 0$, $\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}} = \frac{\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}}{\left\|\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}\right\|_2}$ *for* $\hat{\beta}_{g_{r,\ell}} \neq 0$ *and* $\left\|\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 \leq 1$ *for* $\hat{\beta}_{g_{r,\ell}} = 0$.

Lemma 30. *Take* $\hat{\beta}$ *and* $\hat{a}^{(\ell)}$ *as in the previous lemma. Suppose that*

$$\left\|\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 < 1 \qquad \text{for} \quad \ell = 1, \ldots, J(\hat{\beta})$$

then for any other solution $\tilde{\beta}$ *to* (2.8)*, it is as sparse as* $\hat{\beta}$ *if not more. In other words,*

$$K(\tilde{\beta}) \leq \hat{K}_r.$$

Lemma 31. *(Uniqueness) Under the conditions of the previous lemma, let* $\hat{\mathcal{S}} = \left\{i : \hat{\beta}_i \neq 0\right\}$*. If* $\mathbf{X}_{\hat{\mathcal{S}}}$ *has full column rank (i.e.,* $\text{rank}\,(\mathbf{X}_{\hat{\mathcal{S}}}) = |\hat{\mathcal{S}}|$*) then* $\hat{\beta}$ *is unique.*

Proof. See Appendices A.10, A.11, and A.12. □

A.6 Proof of Theorem 1

We start with introducing notation. **From now on we suppress the dependence on** r **in notation for simplicity**. We denote the group structure $g_\ell = \{1, \cdots, \ell\}$ for $\ell \leq r$ for each $r = 1, \ldots, p$. For any vector $\beta \in \mathbb{R}^r$, we let $\beta_{g_\ell} \in \mathbb{R}^\ell$ be the vector with elements $\{\beta_m : m \leq \ell\}$. We also introduce the weight vector $W^{(\ell)} \in \mathbb{R}^p$ with $\left(W^{(\ell)}\right)_m = w_{\ell m}$ where $w_{\ell m}$ can be defined as in (2.7) or $w_{\ell m} = 1$. Finally recalling from Section 2.4 the definition of $\mathcal{I}$, we denote $\mathcal{S} = \mathcal{I} \cup \{r\} = \{J+1, \ldots, r\}$ and $\mathcal{S}^c = \{1, 2, \ldots, J\}$.

The general idea of the proof depends on the primal-dual witness procedure in Wainwright (2009) and Ravikumar et al. (2011). Considering the original

problem (2.8) for any $r = 2, \ldots, p$, we construct the primal-dual witness solution pairs $\left(\tilde{\beta}, \sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)$ as follows:

(a) Solve the restricted subproblem with the true bandwidth $K = r - 1 - J$:

$$\tilde{\beta} = \underset{\substack{\beta_r > 0 \\ \beta_{\mathcal{S}^c} = 0}}{\arg\min} \left\{ -2 \log \beta_r + \frac{1}{n} \|\mathbf{X}_{1:r}\beta\|_2^2 + \lambda \sum_{\ell=1}^{r-1} \left\| \left(W^{(\ell)} * \beta\right)_{g_\ell} \right\|_2 \right\}.$$

The solution above can be written as

$$\tilde{\beta} = \begin{pmatrix} \mathbf{0}_J \\ \tilde{\gamma} \end{pmatrix},$$

where

$$\tilde{\gamma} = \underset{\gamma \in \mathbb{R}^{K+1}}{\arg\min} \left\{ -2 \log \gamma_{K+1} + \frac{1}{n} \|\mathbf{X}_{\mathcal{S}}\gamma\|_2^2 + \lambda \sum_{\ell=1}^{K} \left\| \left(\tilde{W}^{(\ell)} * \gamma\right)_{g_\ell} \right\|_2 \right\},$$

with

$$\tilde{W}^{(\ell)} = \left(W^{(\ell+J)}\right)_{\mathcal{S}} \quad \Longleftrightarrow \quad \sum_{\ell=1}^{K} \left\| \left(\tilde{W}^{(\ell)} * \gamma\right)_{g_\ell} \right\|_2 = \sum_{\ell=J+1}^{r-1} \sqrt{\sum_{m=J+1}^{r-1} w_{\ell m}^2 \gamma_{m-J}^2}.$$

(b) By Lemma 29, there exist $\tilde{b}^{(\ell)} \in \mathbb{R}^{K+1}$ for $\ell = 1, \ldots, K$, such that $\left(\tilde{b}^{(\ell)}\right)_{g_\ell^c} = 0$ and

$$\left(\tilde{b}^{(\ell)}\right)_{g_{r\ell}} = \frac{\left(\tilde{W}^{(\ell)} * \tilde{\gamma}\right)_{g_\ell}}{\left\| \left(\tilde{W}^{(\ell)} * \tilde{\gamma}\right)_{g_\ell} \right\|_2},$$

satisfying

$$-\frac{2}{\tilde{\gamma}_{K+1}} \mathbf{e}_{K+1} + \frac{2}{n} \mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}} \tilde{\gamma} + \lambda \sum_{\ell=1}^{K} \tilde{W}^{(\ell)} * \tilde{b}^{(\ell)} = 0.$$

(c) For $\ell = J + 1, \ldots, r - 1$, we let

$$\tilde{a}^{(\ell)} = \begin{pmatrix} \mathbf{0}_J \\ \tilde{b}^{(\ell-J)} \end{pmatrix}.$$

Then we have $\left(\tilde{a}^{(\ell)}\right)_{g_\ell^c} = 0$, $\left\| \left(\tilde{a}^{(\ell)}\right)_{g_\ell} \right\|_2 \leq 1$, $\left(\tilde{a}^{(\ell)}\right)_{g_\ell} = \frac{\left(W^{(\ell)} * \tilde{\beta}\right)_{g_\ell}}{\left\| \left(W^{(\ell)} * \tilde{\beta}\right)_{g_\ell} \right\|_2}$ for $\tilde{\beta}_{g_\ell} \neq 0$.

(d) For each $\ell = 1, ..., J$, we choose $\tilde{a}^{(\ell)} \in \mathbb{R}^r$ satisfying

$$\left(\tilde{a}^{(\ell)}\right)_{\ell'} = 0 \quad \text{for any } \ell' \neq \ell \quad \text{and} \quad \left(\tilde{a}^{(\ell)}\right)_{\ell} = -\frac{2}{\lambda w_{\ell\ell}} \left(S\tilde{\beta}\right)_{\ell} = -\frac{2}{n\lambda} \mathbf{X}_{\ell}^T \mathbf{X}_{\mathcal{S}} \tilde{\beta}_{\mathcal{S}}.$$

By construction and the fact that $w_{\ell\ell} = 1$,

$$\lambda \left(W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\ell} = \lambda w_{\ell\ell} \left(\tilde{a}^{(\ell)}\right)_{\ell} = -2 \left(S\tilde{\beta}\right)_{\ell}.$$

By Lemma 29, $\left\{\tilde{a}^{(\ell)}\right\}$ satisfies the optimality condition (A.4):

$$-\frac{2}{\tilde{\beta}_r} \mathbf{e}_r + \frac{2}{n} \mathbf{X}_{1:r}^T \mathbf{X}_{1:r} \tilde{\beta} + \lambda \sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)} = 0 \tag{A.5}$$

(e) Verify the strict dual feasibility condition for $\ell = 1, ..., J$

$$\left| \frac{2}{n\lambda} \mathbf{X}_{\ell}^T \mathbf{X}_{\mathcal{S}} \tilde{\beta}_{\mathcal{S}} \right| = \left| \left(\tilde{a}^{(\ell)}\right)_{\ell} \right| = \left\| \left(\tilde{a}^{(\ell)}\right)_{g_\ell} \right\|_2 < 1. \tag{A.6}$$

At a high level, steps (a) through (d) construct a pair $\left(\tilde{\beta}, \left\{\tilde{a}^{(\ell)}\right\}\right)$ that satisfies the optimality condition (A.4), but the $\left\{\tilde{a}^{(\ell)}\right\}$ is not necessarily guaranteed to be a member of $\partial\left(P(\tilde{\beta})\right)$. Step (e) does more than verifying the necessary conditions for it to belong to $\partial\left(P(\tilde{\beta})\right)$. The strict dual feasibility condition, once verified, ensures the uniqueness of the solution. Note that by construction in Step (b), $\left\{\tilde{a}^{(\ell)}\right\}$ satisfies dual feasibility conditions for $\ell = J+1, ..., r-1$ since $\left\{\tilde{b}^{(\ell)}\right\}$ does, so it remains to verify for $\ell = 1, ..., J$ (see Step (c)).

For each $\ell = 1, ..., J$, by the construction in Step (d), $\left(\tilde{a}^{(\ell)}\right)_{g_\ell^c} = 0$. Note that $\tilde{\beta}_{g_J} = 0$ implies $\tilde{\beta}_{g_\ell} = 0$. Thus, for $\tilde{a}^{(\ell)}$ to satisfy conditions in Lemma 29, it suffices to show (A.6).

If the primal-dual witness procedure succeeds, then by construction, the solution $\tilde{\beta}$, whose support is contained in the support of the true $L_{r\cdot}$, is a solution to (2.8). Moreover, by strict dual feasibility and Lemma 31, we know that $\tilde{\beta}$ is the

unique solution $\hat{\beta}$ to the unconstrained problem (2.8). Therefore, the support of $\hat{\beta}$ is contained in the support of L_r..

In the following we adapt the same proof technique as Wainwright (2009) to show that the primal-dual witness succeeds with high probability, from which we first conclude that $K(\hat{\beta}) \leq K$.

A.6.1 Proof of Property 1 in Theorem 1

Proof. We need to verify the strict dual feasibility (A.6). By (A.5),

$$-\frac{2}{\tilde{\beta}_r} + \frac{2}{n}\mathbf{X}_r^T\mathbf{X}_r\tilde{\beta}_r + \frac{2}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}} = 0, \tag{A.7}$$

$$\frac{2}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\tilde{\beta}_r + \frac{2}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}} + \lambda\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} = 0. \tag{A.8}$$

From (A.8),

$$\tilde{\beta}_{\mathcal{I}} = -\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left[\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\tilde{\beta}_r + \frac{\lambda n}{2}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}\right]. \tag{A.9}$$

Plugging (A.9) back into (A.7) and denoting $\mathbf{C}_{\mathcal{I}} = \mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}$ and $\mathbf{O}_{\mathcal{I}} = \mathbf{I} - \mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^T$ as the orthogonal projection matrix onto the orthogonal complement of the column space of $\mathbf{X}_{\mathcal{I}}$, we have

$$-\frac{2}{\tilde{\beta}_r} + \frac{2}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\tilde{\beta}_r - \lambda\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}} = 0,$$

which implies that

$$\tilde{\beta}_r = \frac{\frac{\lambda}{2}\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}} + \sqrt{\frac{\lambda^2}{4}\left(\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right)^2 + \frac{4}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r}}{\frac{2}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r} \tag{A.10}$$

and that

$$\begin{aligned}\left(\tilde{a}^{(\ell)}\right)_\ell &= -\frac{2}{n\lambda}\mathbf{X}_\ell^T\mathbf{X}_S\tilde{\beta}_S = -\frac{2}{n\lambda}\mathbf{X}_\ell^T\mathbf{X}_r\tilde{\beta}_r - \frac{2}{n\lambda}\mathbf{X}_\ell^T\mathbf{X}_I\tilde{\beta}_I \\ &= -\frac{2}{n\lambda}\mathbf{X}_\ell^T\mathbf{X}_r\tilde{\beta}_r + \frac{2}{n\lambda}\mathbf{X}_\ell^T\mathbf{X}_I\left(\mathbf{X}_I^T\mathbf{X}_I\right)^{-1}\left[\mathbf{X}_I^T\mathbf{X}_r\tilde{\beta}_r + \frac{\lambda n}{2}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_I\right] \\ &= -\frac{2}{n\lambda}\mathbf{X}_\ell^T\left[\mathbf{I} - \mathbf{X}_I\left(\mathbf{X}_I^T\mathbf{X}_I\right)^{-1}\mathbf{X}_I^T\right]\mathbf{X}_r\tilde{\beta}_r + \mathbf{X}_\ell^T\mathbf{X}_I\left(\mathbf{X}_I^T\mathbf{X}_I\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_I \\ &= \mathbf{X}_\ell^T\left[\mathbf{C}_I - \mathbf{O}_I\left(\frac{2}{n\lambda}\mathbf{X}_r\tilde{\beta}_r\right)\right]. \end{aligned} \tag{A.11}$$

Conditioning on $\mathbf{X}_I$, we can decompose $\mathbf{X}_r$ and $\mathbf{X}_\ell$ as

$$\begin{aligned}\mathbf{X}_r^T &= \Sigma_{rI}\left(\Sigma_{II}\right)^{-1}\mathbf{X}_I^T + E_r^T, \\ \mathbf{X}_\ell^T &= \Sigma_{\ell I}\left(\Sigma_{II}\right)^{-1}\mathbf{X}_I^T + E_\ell^T,\end{aligned} \tag{A.12}$$

where $E_r \sim N\left(\mathbf{0}_n, \theta_r^{(r)}\mathbf{I}_{n\times n}\right)$ and $E_\ell \sim N\left(\mathbf{0}_n, \theta_r^{(\ell)}\mathbf{I}_{n\times n}\right)$, and $\theta_r^{(\ell)}$ and $\theta_r^{(r)}$ are defined in Section 4. Then

$$\mathbf{X}_\ell^T\mathbf{O}_I = E_\ell^T\mathbf{O}_I \quad \text{and} \quad \mathbf{O}_I\mathbf{X}_r = \mathbf{O}_I E_r,$$

and from (A.11)

$$\begin{aligned}\left(\tilde{a}^{(\ell)}\right)_\ell &= E_\ell^T\left[\mathbf{C}_I - \mathbf{O}_I\left(\frac{2}{n\lambda}E_r\tilde{\beta}_r\right)\right] + \Sigma_{\ell I}\left(\Sigma_{II}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_I \\ &:= R^{(\ell)} + F^{(\ell)}.\end{aligned} \tag{A.13}$$

We first bound $\max_\ell \left|F^{(\ell)}\right|$. Note that

$$\begin{aligned}\left\|\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_I\right\|_\infty &= \left\|\left(\sum_{\ell=J+1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_I\right\|_\infty = \max_{m\in I}\left|\sum_{\ell=m}^{r-1} w_{\ell m}\left(\tilde{a}^{(\ell)}\right)_m\right| \\ \leq \max_{m\in I}\sum_{\ell=m}^{r-1} w_{\ell m}\left|\left(\tilde{a}^{(\ell)}\right)_m\right| &\leq \max_{m\in I}\sum_{\ell=m}^{r-1}\frac{1}{(\ell-m+1)^2} \leq \sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6},\end{aligned} \tag{A.14}$$

where we used $\left\|\tilde{a}^{(\ell)}\right\|_\infty \le \left\|\tilde{a}^{(\ell)}\right\|_2 \le 1$. Therefore, by Assumption **A3**,

$$\max_{1\le\ell\le J}\left|\Sigma_{\ell\mathcal{I}}\left(\Sigma_{\mathcal{I}\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}\right| \le 1-\alpha.$$

To give a bound on the random quantity $\left|R^{(\ell)}\right|$, we first state a general result that will be used multiple times later in the proof.

Lemma 32. *Consider the term $E_j^T\eta$ where $\eta \in \mathbb{R}^n$ is a random vector depending on $\mathbf{X}_{\mathcal{I}}$ and $\mathbf{X}_r$ and $E_j \sim N\left(\mathbf{0}_n, \theta_r^{(j)}\mathbf{I}_{n\times n}\right)$ for $j = 1,\dots,J,r$. If for some $\bar{Q} \ge 0$*

$$\mathrm{P}\left[\mathrm{Var}\left(E_j^T\eta\middle|\mathbf{X}_{\mathcal{I}},\mathbf{X}_r\right) \ge \bar{Q}\right] \le \bar{p}$$

then for any $a > 0$,

$$\mathrm{P}\left[\left|E_j^T\eta\right| \ge a\right] \le 2\exp\left(-\frac{a^2}{2\bar{Q}}\right) + \bar{p}$$

Proof. Define the event

$$\bar{\mathcal{B}} = \left\{\mathrm{Var}\left(E_j^T\eta\middle|\mathbf{X}_{\mathcal{I}}\right) \ge \bar{Q}\right\}.$$

Now for any a and conditioned on $\mathbf{X}_{\mathcal{I}}$ and $\mathbf{X}_r$,

$$\mathrm{P}\left[E_j^T\eta \ge a\right] \le \mathrm{P}\left[E_j^T\eta \ge a\middle|\bar{\mathcal{B}}^c\right] + \mathrm{P}\left[\bar{\mathcal{B}}\right] \le \mathrm{P}\left[E_j^T\eta \ge a\middle|\bar{\mathcal{B}}^c\right] + \bar{p}.$$

Conditioned on $\bar{\mathcal{B}}^c$, the variance of $E_j^T\eta$ is at most $\bar{Q}$. So by standard Gaussian tail bounds, we have

$$\mathrm{P}\left[E_j^T\eta \ge a\middle|\bar{\mathcal{B}}^c\right] = \mathrm{E}\left[\mathrm{P}\left(E_j^T\eta \ge a\middle|\mathbf{X}_{\mathcal{I}},\mathbf{X}_r\right)\middle|\bar{\mathcal{B}}^c\right] \le \mathrm{E}\left[2\exp\left(-\frac{a^2}{2\bar{Q}}\right)\middle|\bar{\mathcal{B}}^c\right] \le 2\exp\left(-\frac{a^2}{2\bar{Q}}\right).$$

□

Then note that $\mathrm{Var}\left(E_{i\ell}\right) = \theta_r^{(\ell)} \leq \theta_r$ for $i = 1, \ldots, n$. Now conditioned on both $\mathbf{X}_{\mathcal{I}}$ and $\mathbf{X}_r$, $R^{(\ell)}$ is zero-mean with variance at most

$$\begin{aligned}
&\mathrm{Var}\left(R^{(\ell)} \middle| \mathbf{X}_{\mathcal{I}}\right) \\
&\leq \theta_r \left\| \mathbf{C}_{\mathcal{I}} - \mathbf{O}_{\mathcal{I}} \left(\frac{2}{n\lambda} E_r \tilde{\beta}_r \right) \right\|_2^2 = \theta_r \left\{ \mathbf{C}_{\mathcal{I}}^T \mathbf{C}_{\mathcal{I}} + \left\| \mathbf{O}_{\mathcal{I}} \left(\frac{2}{n\lambda} E_r \tilde{\beta}_r \right) \right\|_2^2 \right\} \\
&= \theta_r \left\{ \frac{1}{n} \left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)} \right)_{\mathcal{I}}^T \left(\frac{1}{n} \mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}} \right)^{-1} \left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)} \right)_{\mathcal{I}} + \frac{4\tilde{\beta}_r^2 \left\| \mathbf{O}_{\mathcal{I}} E_r \right\|_2^2}{n^2 \lambda^2} \right\} \\
&:= \theta_r M_n,
\end{aligned}$$

where the first equality holds from Pythagorean identity. The next lemma bounds the random scaling M_n.

Lemma 33. *For $\varepsilon \in \left(0, \frac{1}{2}\right)$, denote*

$$\bar{M}_n(\varepsilon) := \frac{3\kappa^2\pi^2}{2} \frac{K}{n} + \frac{1}{\theta_r^{(r)} (n - K)(1 - \varepsilon)} + \frac{16}{n\lambda^2},$$

then

$$\mathrm{P}\left[M_n \geq \bar{M}_n(\varepsilon) \middle| \mathbf{X}_{\mathcal{I}} \right] \leq 7 \exp\left(-n \min\left\{ \frac{\alpha^2}{3\theta_r^{(r)} \kappa^2 \pi^2 K}, \frac{\varepsilon^2}{4} \left(1 - \frac{K}{n} \right) \right\} \right).$$

Proof. See Appendix A.13. □

Now by Lemma 32 and the union bound,

$$\mathrm{P}\left[\max_{1 \leq \ell \leq J} \left| R^{(\ell)} \right| \geq \alpha \right] \leq 2J \exp\left(-\frac{\alpha^2}{2\theta_r \bar{M}_n(\varepsilon)} \right) + 7 \exp\left(-c_3 n \right), \tag{A.15}$$

for some constant c_3 independent of n and J. By the assumption that $\frac{K}{n} = o(1)$, we have that $\frac{K}{n} \leq 1 - \varepsilon$ for n large enough, thus

$$\bar{M}_n(\varepsilon) \leq \frac{K}{n} \left(\frac{3\kappa^2\pi^2}{2} + \frac{1}{K\theta_r^{(r)} (1 - \varepsilon)^2} + \frac{16}{K\lambda^2} \right) \leq \frac{K}{n} \left(\frac{3\kappa^2\pi^2}{2} + \frac{4}{K\theta_r^{(r)}} + \frac{16}{K\lambda^2} \right).$$

For the exponential term in (A.15) to have faster decaying rate than the J term, we need

$$\frac{n}{K\log J} > \frac{\theta_r}{\alpha^2}\left(3\kappa^2\pi^2 + \frac{8}{K\theta_r^{(r)}} + \frac{32}{K\lambda^2}\right).$$

□

A.6.2 Proof of Property 2 in Theorem 1

Next we study the ℓ_∞ error bound. The following theorem gives an ℓ_∞ error bound of $\tilde{\beta}$.

Proof. Let $\delta = \tilde{\beta} - \beta^* = \tilde{\beta} - \left(L^T\right)_{1:r,r}$ and $\mathcal{W} = SL^T - (L)^{-1}$, then from (A.8) and the fact that L^{-1} is lower-triangular,

$$\begin{aligned}
\delta_{\mathcal{I}} &= -\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left[\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\tilde{\beta}_r + \left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)(L)_{\mathcal{I},r}^T\right] - \frac{n\lambda}{2}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} \\
&= -\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left[\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\left(\delta_r + \beta_r^*\right) + \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)(L)_{\mathcal{I},r}^T\right] - \frac{\lambda}{2}\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} \\
&= -\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\delta_r - \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(SL^T\right)_{\mathcal{I},r} - \frac{\lambda}{2}\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} \\
&= -\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_r\delta_r - \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathcal{W}_{\mathcal{I},r} - \frac{\lambda}{2}\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}.
\end{aligned} \tag{A.16}$$

From (A.7) and the fact that $\left(L^{-1}\right)_{rr} = \frac{1}{L_{rr}}$,

$$\begin{aligned}
&-\frac{1}{\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r\delta_r + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\delta_{\mathcal{I}} + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r\beta_r^* + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}}^* \\
&= -\frac{1}{\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r\delta_r + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\delta_{\mathcal{I}} + \left(SL^T\right)_{rr} \\
&= \left(L^{-1}\right)_{rr} - \frac{1}{\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r\delta_r + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\delta_{\mathcal{I}} + \mathcal{W}_{rr} \\
&= \frac{\delta_r}{L_{rr}\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_r\delta_r + \frac{1}{n}\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\delta_{\mathcal{I}} + \mathcal{W}_{rr} = 0. \qquad \text{(A.17)}
\end{aligned}$$

Plugging (A.16) into (A.17), we have

$$\frac{\delta_r}{L_{rr}\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\delta_r = \mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathcal{W}_{\mathcal{I},r} + \frac{\lambda}{2}\mathbf{X}_r\mathbf{C}_{\mathcal{I}} - \mathcal{W}_{rr},$$

which implies

$$\delta_r = \left(\frac{1}{L_{rr}\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\right)^{-1}\left[\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathcal{W}_{\mathcal{I},r} + \frac{\lambda}{2}\mathbf{X}_r\mathbf{C}_{\mathcal{I}} - \mathcal{W}_{rr}\right].$$

Since $L_{rr} > 0$ and $\tilde{\beta}_r > 0$,

$$\begin{aligned}
|\delta_r| &\le \left|\left(\frac{1}{L_{rr}\tilde{\beta}_r} + \frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\right)^{-1}\right|\left(\left|\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathcal{W}_{\mathcal{I},r}\right| + \left|\frac{\lambda}{2}\mathbf{X}_r\mathbf{C}_{\mathcal{I}}\right| + |\mathcal{W}_{rr}|\right) \\
&\le \left|\left(\frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\right)^{-1}\right|\left(\left|\mathbf{X}_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathcal{W}_{\mathcal{I},r}\right| + \left|\frac{\lambda}{2}\mathbf{X}_r\mathbf{C}_{\mathcal{I}}\right| + |\mathcal{W}_{rr}|\right).
\end{aligned}$$

Now conditioned on $\mathbf{X}_{\mathcal{I}}$, by the decomposition (A.12), $\left(\frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\right)^{-1} = \left(\frac{1}{n}E_r^T\mathbf{O}_{\mathcal{I}}E_r\right)^{-1} = \frac{n}{\|\mathbf{O}_{\mathcal{I}}E_r\|_2^2}$. From Lemma 39, it follows that

$$\mathrm{P}\left[\left(\frac{1}{n}\mathbf{X}_r^T\mathbf{O}_{\mathcal{I}}\mathbf{X}_r\right)^{-1} \ge \frac{1}{\theta_r^{(r)}}\frac{n}{n-K}\frac{1}{1-\varepsilon}\right] \le \exp\left(-\frac{1}{4}(n-K)\varepsilon^2\right).$$

Also, by Lemma 38,

$$\mathrm{P}\left[\left|\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right| \ge 1\right] \le 2\exp\left(-\frac{n\alpha^2}{3\theta_r^{(r)}\kappa^2\pi^2K}\right) + 2\exp\left(-\frac{n}{2}\right).$$

To deal with the rest of terms in (A.17) that involve $\mathcal{W}$, we introduce the following concentration inequality to control its element-wise infinity norm.

Lemma 34. *Let* $\mathcal{W} = SL^T - L^{-1}$. *Under Assumptions* ***A4*** *and* ***A5****, there exist constants* $C_1, C_2, C_3 > 0$ *such that for any* $0 < t \leq 2\kappa$,

$$\mathrm{P}\left[\|\mathcal{W}\|_\infty > t\right] \leq 2p^2 \exp\left(-\frac{C_3 n t^2}{\kappa^2}\right) + 4p \exp\left(-\frac{C_1 n t}{\kappa^2}\right) + 4p \exp\left(-C_2 n t\right).$$

Proof. See Appendix A.14. □

In terms of the event

$$\mathcal{A} = \left\{\|\mathcal{W}\|_\infty \leq \lambda\right\},$$

Lemma 34 states that

$$\mathrm{P}\left[\mathcal{A}^c\right] \leq 2p^2 \exp\left(-\frac{C_3 n \lambda^2}{\kappa^2}\right) + 4p \exp\left(-\frac{C_1 n \lambda}{\kappa^2}\right) + 4p \exp\left(-C_2 n \lambda\right).$$

The next lemma shows that, on the event $\mathcal{A}$ and with the assumption that $\frac{\lambda^2}{n} = o(1)$, the term $\left|\mathbf{X}_r^T \mathbf{X}_I \left(\mathbf{X}_I^T \mathbf{X}_I\right)^{-1} \mathcal{W}_{I,r}\right|$ can be bounded by λ with high probability.

Lemma 35. *Using the general weigthing scheme* (2.7)*, we have*

$$\mathrm{P}\left[\left|\mathbf{X}_r^T \mathbf{X}_I \left(\mathbf{X}_I^T \mathbf{X}_I\right)^{-1} \mathcal{W}_{I,r}\right| \geq \lambda \middle| \mathcal{A}\right] \leq 2 \exp\left(-\frac{2n\alpha^2}{9\theta_r^{(r)} \kappa^2 K \lambda^2}\right) + 2 \exp\left(-\frac{n}{2}\right).$$

Proof. Recall that by conditioning on $\mathbf{X}_I$, the decomposition (A.12) gives

$$\mathbf{X}_r^T \mathbf{X}_I \left(\mathbf{X}_I^T \mathbf{X}_I\right)^{-1} \mathcal{W}_{I,r} = \Sigma_{rI} \left(\Sigma_{II}\right)^{-1} \mathcal{W}_{I,r} + E_r^T \mathbf{X}_I \left(\mathbf{X}_I^T \mathbf{X}_I\right)^{-1} \mathcal{W}_{I,r}.$$

On the event $\mathcal{A}$, by **A3** and (A.14),

$$\left|\Sigma_{rI} \left(\Sigma_{II}\right)^{-1} \mathcal{W}_{I,r}\right| \leq \left\|\Sigma_{rI} \left(\Sigma_{II}\right)^{-1}\right\|_\infty \left\|\mathcal{W}_{I,r}\right\|_\infty \leq \lambda.$$

Note that $\mathrm{Var}(E_{ir}) = \theta_r^{(r)}$ for $i = 1, \ldots, n$. Let $B^{(r)} := E_r^T \mathbf{X}_{\mathcal{I}} \left(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} \mathcal{W}_{\mathcal{I},r}$, then $B^{(r)}$ has mean zero and variance at most

$$\mathrm{Var}\left(B^{(r)} \middle| \mathbf{X}_{\mathcal{I}}\right) = \frac{\theta_r^{(r)}}{n} \mathcal{W}_{\mathcal{I},r}^T \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} \mathcal{W}_{\mathcal{I},r} \leq \frac{9\theta_r^{(r)} \kappa^2 K \lambda^2}{n},$$

with probability greater than $1 - 2\exp\left(\frac{n}{2}\right)$. The result follows from Lemma 32. □

Putting everything together and choosing the tuning parameter from (2.13), with a union bound argument and some algebra, we have shown that conditioned on $\mathbf{X}_{\mathcal{I}}$,

$$\begin{aligned}
&\mathrm{P}\left[|\delta_r| \geq \frac{1}{\theta_r^{(r)}} \frac{n}{n-K} \frac{1}{1-\varepsilon} \frac{5}{2}\lambda\right] \leq \mathrm{P}\left[|\delta_r| \geq \frac{5}{2\theta_r^{(r)}}\lambda\right] \leq \mathrm{P}\left[|\delta_r| \geq \frac{5}{2\theta_r^{(r)}}\lambda \middle| \mathcal{A}\right] + \mathrm{P}\left[\mathcal{A}^c\right] \\
&\leq \exp\left(-\frac{1}{4n}\left(1 - \frac{K}{n}\right)\varepsilon^2\right) + 2\exp\left(-\frac{n\alpha^2}{3\theta_r \kappa^2 \pi^2 K}\right) + 2\exp\left(-\frac{2n\alpha^2}{9\theta_r \kappa^2 K \lambda^2}\right) + 4\exp\left(-\frac{n}{2}\right) \\
&+ 2p^2 \exp\left(-\frac{C_3 n \lambda^2}{\kappa^2}\right) + 4p\exp\left(-\frac{C_1 n \lambda}{\kappa^2}\right) + 4p\exp\left(-C_2 n\lambda\right) \\
&\leq c_4 \exp\left(-c_5 n\right) + \frac{c_6}{p}, \qquad \text{(A.18)}
\end{aligned}$$

for some constants $c_4, c_5, c_6, x > 0$ that do not depend on n and p.

We now consider a bound for $\delta_{\mathcal{I}}$. Recall from (A.16) that

$$\delta_{\mathcal{I}} = F_1 + F_2$$

where

$$\begin{aligned}
F_1 &= -\left(\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} \mathbf{X}_{\mathcal{I}}^T \mathbf{X}_r \delta_r, \\
F_2 &= -\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} \left(\mathcal{W}_{\mathcal{I},r} + \frac{\lambda}{2}\mathbf{D}\right) \quad \text{with} \quad \mathbf{D} = \left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}.
\end{aligned}$$

An ℓ_∞ bound of F_2 is given by

$$\|F_2\|_\infty \leq \left\|\left(\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T \mathbf{X}_{\mathcal{I}}\right)^{-1} - (\Sigma_{\mathcal{I}\mathcal{I}})^{-1}\right)\left(\mathcal{W}_{\mathcal{I},r} + \frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty + \left\|(\Sigma_{\mathcal{I}\mathcal{I}})^{-1}\left(\mathcal{W}_{\mathcal{I},r} + \frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty. \qquad \text{(A.19)}$$

On the event $\mathcal{A}$, by (A.14),

$$\begin{aligned}\left\|(\Sigma_{II})^{-1}\left(\mathcal{W}_{I,r}+\frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty &\le \left|\!\left|\!\left|(\Sigma_{II})^{-1}\right|\!\right|\!\right|_\infty \left(\left\|\mathcal{W}_{I,r}\right\|_\infty+\frac{\lambda}{2}\|\mathbf{D}\|_\infty\right)\\ &\le \left|\!\left|\!\left|(\Sigma_{II})^{-1}\right|\!\right|\!\right|_\infty \left(1+\frac{\pi^2}{12}\right)\lambda \le 2\lambda\left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty^2\end{aligned}$$

To deal with the first term in (A.19), note that $X_I = W_I(\Sigma_{II})^{1/2}$, where $W_I \in \mathbb{R}^{n\times K}$ is a standard Gaussian random matrix, i.e., $(W_I)_{ij} \sim N(0,1)$. Thus we can write it as

$$\left\|(\Sigma_{II})^{-1/2}\left[\left(\frac{1}{n}W_I^TW_I\right)^{-1}-\mathbf{I}_K\right](\Sigma_{II})^{-1/2}\left(\mathcal{W}_{I,r}+\frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty \le \left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty G,$$

where

$$G=\left\|\left[\left(\frac{1}{n}W_I^TW_I\right)^{-1}-\mathbf{I}_K\right](\Sigma_{II})^{-1/2}\left(\mathcal{W}_{I,r}+\frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty.$$

By Lemma 5 in Wainwright (2009), we have, for some constant $c_7 > 0$.

$$\mathrm{P}\left[G \ge \left\|(\Sigma_{II})^{-1/2}\left(\mathcal{W}_{I,r}+\frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty \middle| \mathbf{X}_I\right] \le 4\exp\left(-c_7\min\{K,\log J\}\right)$$

Note that conditioning on $\mathcal{A}$, $\left\|(\Sigma_{II})^{-1/2}\left(\mathcal{W}_{I,r}+\frac{\lambda}{2}\mathbf{D}\right)\right\|_\infty$ is upper bounded by $2\lambda\left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty$. Thus,

$$\mathrm{P}\left[G \ge 2\lambda\left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty^2 \middle| \mathcal{A}\right] \le 4\exp\left(-c_7\min\{K,\log J\}\right),$$

and

$$\begin{aligned}&\mathrm{P}\left[\|F_2\|_\infty \ge 4\lambda\left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty^2\right] \le \mathrm{P}\left[\|F_2\|_\infty \ge 4\lambda\left|\!\left|\!\left|(\Sigma_{II})^{-1/2}\right|\!\right|\!\right|_\infty^2 \middle| \mathcal{A}\right] + \mathrm{P}\left[\mathcal{A}^c\right]\\ &\le 4\exp\left(-c_7\min\{K,\log J\}\right)+\frac{c_6}{p}. \qquad \text{(A.20)}\end{aligned}$$

Turning to F_1, conditioned on $\mathbf{X}_I$, by decomposition (A.12), we have that

$$\|F_1\|_\infty \le \left\|(\Sigma_{II})^{-1}\Sigma_{Ir}\right\|_\infty|\delta_r| + \left\|\left(\mathbf{X}_I^T\mathbf{X}_I\right)^{-1}\mathbf{X}_I^TE_r\delta_r\right\|_\infty.$$

By (A.18) and **A3**,

$$\mathrm{P}\left[\left\|\left(\Sigma_{\mathcal{I}\mathcal{I}}\right)^{-1}\Sigma_{\mathcal{I}r}\right\|_\infty|\delta_r|\geq\frac{5}{2\theta_r^{(r)}}\lambda\right]\leq c_4\exp\left(-c_5n\right)+\frac{c_6}{p}.$$

Consider each coordinate $j\in\mathcal{I}$ of the random term whose variance is bounded by

$$\mathrm{Var}\left[\mathbf{e}_j^T\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^TE_r\delta_r\middle|\mathbf{X}_{\mathcal{I}}\right]\leq\theta_r\left\|\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\right\|_2\frac{\delta_r^2}{n}.$$

By Lemma 37 and (A.18),

$$\mathrm{P}\left[\mathrm{Var}\left[\mathbf{e}_j^T\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^TE_r\delta_r\middle|\mathbf{X}_{\mathcal{I}}\right]\geq\frac{235}{4}\frac{\kappa^2}{\theta_r}\frac{\lambda^2}{n}\right]\leq2\exp\left(-\frac{n}{2}\right)+c_4\exp\left(-c_5n\right)+\frac{c_6}{p}.$$

Thus by Lemma 32,

$$\begin{aligned}\mathrm{P}\left[\left\|\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathbf{X}_{\mathcal{I}}^TE_r\delta_r\right\|_\infty\geq\frac{5}{2\theta_r^{(r)}}\lambda\right]\leq&2\exp\left(-\frac{n}{18\theta_r\kappa^2}\right)+2\exp\left(-\frac{n}{2}\right)\\&+c_4\exp\left(-c_5n\right)+\frac{c_6}{p},\end{aligned}$$

and

$$\mathrm{P}\left[\|F_1\|_\infty\geq\frac{5}{\theta_r^{(r)}}\lambda\right]\leq2\exp\left(-\frac{n}{18\theta_r\kappa^2}\right)+2\exp\left(-\frac{n}{2}\right)+c_4\exp\left(-c_5n\right)+\frac{c_6}{p}.$$

Combining with (A.18) and (A.20), we have

$$\mathrm{P}\left[\|\delta\|_\infty\geq4\lambda\left\|\left(\Sigma_{\mathcal{I}\mathcal{I}}\right)^{-1/2}\right\|_\infty^2+\frac{5}{\theta_r^{(r)}}\lambda\right]\leq c_8\exp\left(-c_9n\right)+2\frac{c_6}{p}+4\exp\left(-c_7\min\left\{K,\log J\right\}\right),$$

for some constants $c_8,c_9>0$ that do not depend on n and J. □

A.6.3 Proof of Property 3 in Theorem 1

Finally we establish a $\beta_{\min}$ condition, which, combined with the ℓ_∞ rate, gives the other direction of the support recovery, i.e., $K(\hat{\beta})\geq K$.

By the triangle inequality

$$\left|\tilde{\beta}_j\right| \geq \left|\beta_j\right| - \left|\tilde{\beta}_j - \beta_j\right|.$$

So if we have

$$\max_{j \geq J+1} \left\{\left|\beta_j\right| - \left|\tilde{\beta}_j - \beta_j\right|\right\} > 0,$$

then $K(\tilde{\beta}) \geq K$.

A.7 Proof of Theorem 3

Proof. The overall proof techniques are the same as the proof of Theorem 1. The first part of the theorem holds if $\max_{2 \leq r \leq p} \max_{1 \leq \ell \leq J_r} |\tilde{a}^{(r\ell)}| < 1$. Now for each $r = 2, \ldots, p$ we proceed with the same primal-dual witness procedure and end up with the same decomposition (A.13).

Assumption **A3** ensures that $\max_{2 \leq r \leq p} \max_{1 \leq \ell \leq J_r} |F^{(r\ell)}| \leq 1 - \alpha$. Following the same line of proof to deal with random term $R^{(r\ell)}$, we have that $R^{(r\ell)}$ is zero-mean Gaussian with conditional variance bounded above by the scaling

$$\begin{aligned}\theta_r \bar{M}_n^{(r)}(\varepsilon) &= \frac{3\kappa^2\pi^2\theta_r}{2}\frac{K_r^*}{n} + \frac{\theta_r}{\theta_r^{(r)}}\frac{1}{(n - K_r^*)(1-\varepsilon)} + \frac{16\theta_r}{n\lambda^2} \\ &\leq \frac{3\kappa^2\pi^2\theta_r}{2}\left(\frac{K}{n} + \frac{\kappa^2}{n\theta_r^{(r)}(1-\varepsilon)^2} + \frac{16}{n\lambda^2}\right),\end{aligned}$$

for $\varepsilon \in \left(0, \frac{1}{2}\right)$ with high probability, where we use the fact that $K = o(n)$ implies that $\frac{K}{n} \leq \varepsilon$ for n large. And

$$\mathrm{P}\left[\left|R^{(r\ell)}\right| \geq \alpha\right] \leq 2\exp\left(-\frac{\alpha^2}{2\theta_r \bar{M}_n^{(r)}(\varepsilon)}\right) + 7\exp\left(-c_3 n\right).$$

Thus,

$$\begin{aligned}\mathrm{P}\left[\max_{2\le r\le p}\max_{1\le \ell\le J_r}\left|R^{(r\ell)}\right|\ge\alpha\right] &\le 2\sum_{r=2}^{p} J_r \exp\left(-\frac{\alpha^2}{2\theta_r \bar{M}_n^{(r)}(\varepsilon)}\right) + 7\sum_{r=2}^{p} J_r \exp\left(-c_3 n\right)\\ &\le p^2 \exp\left(-\frac{\alpha^2}{3\kappa^2\pi^2\theta\frac{K}{n}+\frac{8\theta\kappa^2}{n}+\frac{32\theta}{n\lambda^2}}\right)+\frac{7}{2}p^2\exp\left(-c_3 n\right).\end{aligned}$$

For the exponential term to decay faster than p^2, we need

$$\frac{n}{\log p} > \max\left\{\frac{2}{\alpha^2}\left(3\kappa^2\pi^2\theta K + 8\kappa^2\theta + \frac{32\theta}{\lambda^2}\right), \frac{2}{c_3}\right\}.$$

□

A.8 Proof of Theorem 4

Lemma 36. *Using the notation and conditions in Theorem 4, the following deviation bounds hold with high probability:*

$$\begin{aligned}\left|\!\left|\!\left|\hat{L}-L\right|\!\right|\!\right|_\infty &\le \zeta_\Gamma (K+1)\sqrt{\frac{\log p}{n}},\\ \left|\!\left|\!\left|\hat{L}-L\right|\!\right|\!\right|_1 &\le \zeta_\Gamma (K+1)\sqrt{\frac{\log p}{n}},\\ \left|\!\left|\!\left|\hat{L}-L\right|\!\right|\!\right|_2 &\le \zeta_\Gamma (K+1)\sqrt{\frac{\log p}{n}},\\ \left\|\hat{L}-L\right\|_F &\le \zeta_\Gamma\sqrt{\frac{(s+p)\log p}{n}}.\end{aligned}$$

Proof. By Theorem 3, with high probability, the support of $\hat{L}$ is contained in the true support and

$$\left\|\hat{L}-L\right\|_\infty \le \zeta_\Gamma\sqrt{\frac{\log p}{n}}.$$

Note that

$$\left|\!\left|\!\left|\hat{L}-L\right|\!\right|\!\right|_\infty = \max_{2\le r\le p}\sum_{c=1}^{r}\left|\hat{L}_{rc}-L_{rc}\right| \le \max_{2\le r\le p}(K_r+1)\left\|\hat{L}-L\right\|_\infty \le (K+1)\left\|\hat{L}-L\right\|_\infty.$$

Denote $D = \max_{1\le c\le p-1} D_c$ where $D_c = |\{r = c, \ldots, p : L_{rc} \neq 0\}|$. Observing that $D \le K$, we have

$$\begin{aligned}\left\|\left\|\hat{L} - L\right\|\right\|_1 &= \max_{1\le c\le p-1} \sum_{r=1}^{c} \left|\hat{L}_{rc} - L_{rc}\right| \le \max_{1\le c\le p-1} (D_c + 1) \left\|\hat{L} - L\right\|_\infty \\ &\le (D+1) \left\|\hat{L} - L\right\|_\infty \le (K+1) \left\|\hat{L} - L\right\|_\infty .\end{aligned}$$

By Hölder's inequality

$$\left\|\left\|\hat{L} - L\right\|\right\|_2 \le \sqrt{\left\|\left\|\hat{L} - L\right\|\right\|_1 \left\|\left\|\hat{L} - L\right\|\right\|_\infty}.$$

Finally for Frobenius norm,

$$\left\|\hat{L} - L\right\|_F^2 = \sum_{r=2}^{p} \sum_{c=J_r+1}^{r} \left(\hat{L}_{rc} - L_{rc}\right)^2 \le \sum_{r=2}^{p} \sum_{c=J_r+1}^{r} \left\|\hat{L} - L\right\|_\infty^2 \le \zeta_\Gamma^2 \left(\sum_r K_r + p\right) \frac{\log p}{n}.$$

□

of Theorem 4. First note that

$$\begin{aligned}\hat{L}^T\hat{L} - L^T L &= \left(\hat{L} - L\right)^T \left(\hat{L} - L\right) + \hat{L}^T L + L^T \hat{L} - 2L^T L \\ &= \left(\hat{L} - L\right)^T \left(\hat{L} - L\right) + \left(\hat{L} - L\right)^T L + L^T \left(\hat{L} - L\right).\end{aligned}$$

Thus,

$$\begin{aligned}\left\|\hat{L}^T\hat{L} - L^T L\right\|_\infty &\le \left\|\left\|\hat{L} - L\right\|\right\|_\infty \left\|\hat{L} - L\right\|_\infty + 2\|L\|_\infty \left\|\hat{L} - L\right\|_\infty , \\ \left\|\left\|\hat{L}^T\hat{L} - L^T L\right\|\right\|_1 &= \left\|\left\|\hat{L}^T\hat{L} - L^T L\right\|\right\|_\infty \le 2\|L\|_\infty \left\|\left\|\hat{L} - L\right\|\right\|_\infty + \left\|\left\|\hat{L} - L\right\|\right\|_\infty^2.\end{aligned}$$

By Hölder's inequality

$$\left\|\left\|\hat{L}^T\hat{L} - L^T L\right\|\right\|_2 \le \sqrt{\left\|\left\|\hat{L}^T\hat{L} - L^T L\right\|\right\|_1 \left\|\left\|\hat{L}^T\hat{L} - L^T L\right\|\right\|_\infty}.$$

Finally, for Frobenius norm, observe that

$$\begin{aligned}\left\|L^T\left(\hat{L} - L\right)\right\|_F &= \left\|\mathrm{vec}\left(L^T\left(\hat{L} - L\right)\right)\right\|_2 = \left\|\left(I_p \otimes L^T\right)\mathrm{vec}\left(\hat{L} - L\right)\right\|_2 \\ &\le \left\|I_p \otimes L^T\right\|_2 \left\|\hat{L} - L\right\|_F = \|L\|_2 \left\|\hat{L} - L\right\|_F .\end{aligned}$$

Applying the same strategy to $\left\|\left(\hat{L}-L\right)\left(\hat{L}-L\right)\right\|_F$, we have

$$\left\|\hat{L}^T\hat{L}-L^TL\right\|_F \leq \left(\left\|\hat{L}-L\right\|_2 + 2\|L\|_2\right)\left\|\hat{L}-L\right\|_F,$$

then the results follow from Corollary 36. □

A.9 Proof of Theorem 6

Proof. We adapt the proof technique of Rothman et al. (2008). Let

$$\begin{aligned} G(\Delta) = &-2\log\det(L+\Delta) + \operatorname{tr}\left(S\,(L+\Delta)^T(L+\Delta)\right) + \lambda\,\|(\Delta+L)\|_{2,1}^* \\ &+ 2\log\det L - \operatorname{tr}\left(S\,L^TL\right) - \lambda\,\|L\|_{2,1}^*, \end{aligned} \tag{A.21}$$

where L is the inverse of the Cholesky factor of the true covariance matrix, and the penalty is defined above as

$$\|L\|_{2,1}^* = \sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 L_{rm}^2}.$$

Since the estimator $\hat{L}$ is defined as

$$\hat{L} = \underset{L_{jk}=0:\,j<k}{\arg\min}\left\{-2\log\det L + \operatorname{tr}\left(S\,L^TL\right) + \lambda\,\|L\|_{2,1}^*\right\},$$

it follows that $G(\Delta)$ is minimized at $\hat{\Delta} = \hat{L} - L$. Consider the value of $G(\Delta)$ on the set defined as

$$\Theta_n(M) = \left\{\Delta:\ \Delta_{jk} = 0 \text{ for all } k > j,\quad (\Delta+L)_{jj} > 0 \text{ for all } j\ , \|\Delta\|_F = Mr_n\right\},$$

where $M > 0$ and

$$r_n = \sqrt{\frac{\left(\sum_{r=2}^{p} K_r + p\right)\log p}{n}}.$$

The assumed scaling implies that $r_n \to 0$. We aim at showing that $\inf\{G(\Delta) : \Delta \in \Theta_n(M)\} > 0$. If it holds, then the convexity of $G(\Delta)$ and the fact that $G(\hat{\Delta}) \leq G(\mathbf{0}) = 0$ implies

$$\|\hat{\Delta}\|_F = \|\hat{L} - L\|_F \leq M r_n.$$

We start with analyzing the logarithm terms in (A.21). First let $f(t) = \log\det(L + t\Delta)$. Using a Taylor expansion of $f(t)$ at $t = 0$ with $f'(t) = \mathrm{tr}[(L + t\Delta)^{-1}\Delta]$ and $f''(t) = -\mathrm{vec}\,\Delta^T (L + t\Delta)^{-1} \otimes (L + t\Delta)^{-1}\,\mathrm{vec}\,\Delta$, we have

$$\begin{aligned}
&\log\det(L + \Delta) - \log\det(L) \\
&= \mathrm{tr}(L^{-1}\Delta) - (\mathrm{vec}\,\Delta)^T \left[\int_0^1 (1 - v)(L + v\Delta)^{-1} \otimes (L + v\Delta)^{-1} dv\right] (\mathrm{vec}\,\Delta).
\end{aligned}$$

The trace term in (A.21) can be written as

$$\begin{aligned}
\mathrm{tr}\left(S\,(L + \Delta)^T\,(L + \Delta)\right) - \mathrm{tr}\left(S L^T L\right) &= \mathrm{tr}\left(S L^T \Delta + S\Delta^T L + S\Delta^T \Delta\right) \\
&= 2\,\mathrm{tr}\left(S L^T \Delta\right) + \mathrm{tr}\left(S\Delta^T \Delta\right) \\
&\geq 2\,\mathrm{tr}\left(S L^T \Delta\right),
\end{aligned}$$

where the last inequality comes from the fact that the sample covariance matrix S is positive semidefinite. Combining these with (A.21) gives

$$\begin{aligned}
G(\Delta) \geq& 2(\mathrm{vec}\,\Delta)^T \left[\int_0^1 (1 - v)(L + v\Delta)^{-1} \otimes (L + v\Delta)^{-1} dv\right] (\mathrm{vec}\,\Delta) \\
&+ 2\,\mathrm{tr}[(S L^T - L^{-1})\Delta] + \lambda\left(\|L + \Delta\|_{2,1}^* - \|L\|_{2,1}^*\right) \\
\equiv& (a) + (b) + (c). \qquad \text{(A.22)}
\end{aligned}$$

The integral term (a) above has a positive lower bound. Recalling that $\sigma_{\min}(M) = \min_{\|x\|=1} x^T M x$ is a concave function of M (the minimum of linear func-

tions of M is concave), we have

$$\begin{aligned}
(a) &= 2\|\operatorname{vec}\Delta\|^2 \frac{\operatorname{vec}\Delta^T}{\|\operatorname{vec}\Delta\|}\left[\int_0^1 (1-v)(L+v\Delta)^{-1}\otimes(L+v\Delta)^{-1}dv\right]\frac{\operatorname{vec}\Delta}{\|\operatorname{vec}\Delta\|} \\
&\geq 2\|\Delta\|_F^2\sigma_{\min}\left[\int_0^1 (1-v)(L+v\Delta)^{-1}\otimes(L+v\Delta)^{-1}dv\right] \\
&\geq 2\|\Delta\|_F^2\left[\int_0^1 (1-v)\sigma_{\min}\left((L+v\Delta)^{-1}\otimes(L+v\Delta)^{-1}\right)dv\right] \\
&\geq 2\|\Delta\|_F^2\int_0^1 (1-v)\sigma_{\min}^2(L+v\Delta)^{-1}dv \\
&\geq \|\Delta\|_F^2 \min_{0\leq v\leq 1}\sigma_{\min}^2(L+v\Delta)^{-1} \\
&\geq \|\Delta\|_F^2 \min\left\{\sigma_{\min}^2(L+\tilde{\Delta})^{-1} :\ \|\tilde{\Delta}\|_F \leq Mr_n\right\}. \qquad \text{(A.23)}
\end{aligned}$$

The second inequality uses Jenson's inequality of the concave function $\sigma_{\min}(\cdot)$, and the third inequality uses the fact that $\sigma_{\min}(A\otimes A) = \sigma_{\min}(A)^2$ for any positive (semi)definite matrix A. Using triangle inequality on the matrix operator norm, we have

$$\sigma_{\min}^2(L+\tilde{\Delta})^{-1} = \sigma_{\max}^{-2}(L+\tilde{\Delta}) \geq \left(\|L\|_2 + \left\|\tilde{\Delta}\right\|_2\right)^{-2} \geq \frac{1}{2\|L\|_2^2} \geq \frac{\kappa^2}{2},$$

where the second inequality holds with high probability since $\left\|\tilde{\Delta}\right\|_2 \leq \|\tilde{\Delta}\|_F \leq Mr_n \leq \|L\|_2$ as $r_n \to 0$ and the last inequality follows from Assumption **A4**. This gives the lower bound for the first term in (A.22):

$$(a) \geq \frac{1}{2}\kappa^2\|\Delta\|_F^2 = \frac{1}{2}\kappa^2 M^2 r_n^2. \qquad \text{(A.24)}$$

To deal with (b), we start by recalling some notation. We let $\mathcal{S} = \left\{(r,j) : L_{rj} \neq 0\right\}$ denote the support of L, and $s = \sum_{r=2}^p K_r$ be the number of non-zero off-diagonal elements. We also define

$$\|L\|_{2,1} = \sum_{r=2}^p\sum_{\ell=1}^{r-1} w_{\ell\ell}|L_{r\ell}| = \sum_{r=2}^p\sum_{\ell=1}^{r-1}|L_{r\ell}|,$$

where the last equality holds since $w_{\ell\ell} = 1$ by (2.7). Then, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\mathrm{tr}[(SL^T - L^{-1})\Delta]\right| &= \left|\sum_{r=1}^{p}\sum_{j=1}^{r}\left(SL^T - L^{-1}\right)_{rj}\Delta_{rj}\right| \\
&\leq \left|\sum_{r=1}^{p}\sum_{j\in\mathcal{I}_r}(SL^T - L^{-1})_{rj}\Delta_{rj}\right| + \left|\sum_{r=1}^{p}\sum_{j\notin\mathcal{I}_r}(SL^T - L^{-1})_{rj}\Delta_{rj}\right| \\
&\leq \sqrt{s+p}\left\|SL^T - L^{-1}\right\|_\infty \|\Delta_{\mathcal{S}}\|_F + \left\|SL^T - L^{-1}\right\|_\infty \|\Delta_{\mathcal{S}^c}\|_{2,1} \\
&\leq C_1\sqrt{s+p}\sqrt{\frac{\log p}{n}}\,\|\Delta_{\mathcal{S}}\|_F + C_1\sqrt{\frac{\log p}{n}}\,\|\Delta_{\mathcal{S}^c}\|_{2,1}\,, \qquad \text{(A.25)}
\end{aligned}$$

where the last inequality comes from Lemma 34 with probability tending to 1. To bound the penalty terms, we note that

$$\begin{aligned}
&\|L+\Delta\|_{2,1}^* - \|L\|_{2,1}^* \\
&= \sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 (L_{rm}+\Delta_{rm})^2} - \|L_{\mathcal{S}}\|_{2,1}^* \\
&= \sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m:(r,m)\in\mathcal{S}} w_{\ell m}^2 (L_{rm}+\Delta_{rm})^2 + \sum_{m:(r,m)\notin\mathcal{S}} w_{\ell m}^2 (L_{rm}+\Delta_{rm})^2} - \|L_{\mathcal{S}}\|_{2,1}^* \\
&\geq \sum_{r=2}^{p}\sum_{\ell=1}^{r-1}\sqrt{\sum_{m:(r,m)\in\mathcal{S}} w_{\ell m}^2 (L_{rm}+\Delta_{rm})^2} + \sum_{r=2}^{p}\sum_{\ell:(r,\ell)\notin\mathcal{S}} |L_{r\ell}+\Delta_{r\ell}| - \|L_{\mathcal{S}}\|_{2,1}^* \\
&= \|L_{\mathcal{S}}+\Delta_{\mathcal{S}}\|_{2,1}^* + \|L_{\mathcal{S}^c}+\Delta_{\mathcal{S}^c}\|_{2,1} - \|L_{\mathcal{S}}\|_{2,1}^* \\
&= \|L_{\mathcal{S}}+\Delta_{\mathcal{S}}\|_{2,1}^* + \|\Delta_{\mathcal{S}^c}\|_{2,1} - \|L_{\mathcal{S}}\|_{2,1}^* \\
&\geq \|\Delta_{\mathcal{S}^c}\|_{2,1} - \|\Delta_{\mathcal{S}}\|_{2,1}^*\,,
\end{aligned}$$

where the last inequality comes from triangle inequality. To give an upper bound on $\|L_{\mathcal{S}}\|_{2,1}^*$, we observe that $2\lambda b \leq a\lambda^2 + b^2/a$ holds for any $a > 0$, and

obtain

$$\begin{aligned}
2\lambda \|\Delta_{\mathcal{S}}\|_{2,1}^* &= \sum_{r=2}^{p} 2\lambda \sum_{\ell=J_r+1}^{r-1} \sqrt{\sum_{m=J_r+1}^{\ell} w_{\ell m}^2 \Delta_{rm}^2} \\
&\leq \left(\sum_{r=2}^{p} K_r\right) \lambda^2 a + \sum_{r=2}^{p} \sum_{\ell=J_r+1}^{r-1} \sum_{m=J_r+1}^{\ell} w_{\ell m}^2 \Delta_{rm}^2 / a \\
&= \left(\sum_{r=2}^{p} K_r\right) \lambda^2 a + \sum_{r=2}^{p} \sum_{m=J_r+1}^{r-1} \left(\sum_{\ell=m}^{r-1} w_{\ell m}^2\right) \Delta_{rm}^2 / a.
\end{aligned}$$

Now let

$$\begin{aligned}
a &= \frac{4}{\kappa^2} \max_r \max_{J_r+1 \leq m \leq r-1} \sum_{\ell=m}^{r-1} w_{\ell m}^2 \\
&= \frac{4}{\kappa^2} \max_r \max_{J_r+1 \leq m \leq r-1} \sum_{\ell=m}^{r-1} \frac{1}{(\ell - m + 1)^4} \leq \sum_{k=1}^{\infty} \frac{4}{k^4 \kappa^2} \leq \frac{C_2}{\kappa^2},
\end{aligned}$$

for some constant $C_2 > 0$, it follows that

$$\lambda \|\Delta_{\mathcal{S}}\|_{2,1}^* \leq \frac{C_2}{\kappa^2} s\lambda^2 + \|\Delta_{\mathcal{S}}\|_F^2 \frac{\kappa^2}{4} \leq \frac{C_2}{\kappa^2} s\lambda^2 + \|\Delta\|_F^2 \frac{\kappa^2}{4}.$$

Therefore,

$$\lambda \left(\|L + \Delta\|_{2,1}^* - \|L\|_{2,1}^*\right) \geq \lambda \|\Delta_{\mathcal{S}^c}\|_{2,1} - \frac{C_2}{\kappa^2} s\lambda^2 - \frac{\kappa^2}{4} \|\Delta\|_F^2 . \tag{A.26}$$

Finally, combining (A.24), (A.25), and (A.26), we have

$$G(\Delta) \geq \frac{\kappa^2}{4} \|\Delta\|_F^2 - C_1 \sqrt{\frac{(s+p)\log p}{n}} \|\Delta\|_F + \left(\lambda - C_1 \sqrt{\frac{\log p}{n}}\right) \|\Delta_{\mathcal{S}^c}\|_{2,1} - \frac{C_2}{\kappa^2} s\lambda^2.$$

For any $\varepsilon < 1$, choose

$$\lambda = \frac{C_1}{\varepsilon} \sqrt{\frac{\log p}{n}}.$$

Since $\|\Delta\|_F = M r_n$, we have

$$\begin{aligned}
G(\Delta) &\geq \frac{\kappa^2}{4} M^2 r_n^2 - C_1 M r_n^2 + C_1 \sqrt{\frac{\log p}{n}} \left(\frac{1}{\varepsilon} - 1\right) \|\Delta_{\mathcal{S}^c}\|_{2,1} - \frac{C_2 C_1^2}{\kappa^2 \varepsilon^2} \frac{s \log p}{n} \\
&\geq \left(\frac{\kappa^2}{4} M^2 - C_1 M - \frac{C_2 C_1^2}{\kappa^2 \varepsilon^2}\right) r_n^2 > 0,
\end{aligned}$$

for M sufficiently large. □

A.10 Proof of Lemma 29

Proof. Denote

$$\begin{aligned}
&\mathcal{L}\left(\tau, z, \beta; \nu, \phi, a^{(\ell)}\right) \\
=& -2\log\tau + \frac{1}{n}\|z\|_2^2 + \nu(\tau - \beta_r) + \frac{1}{n}\langle\phi, z - \mathbf{X}_{1:r}\beta\rangle + \lambda\sum_{\ell=1}^{r-1}\left\langle W^{(\ell)} * a^{(\ell)}, \beta\right\rangle.
\end{aligned}$$

Then the primal (2.8) can be written equivalently as

$$\min_{\tau,z,\beta}\left\{\max_{\nu,\phi,a^{(\ell)}}\left\{\mathcal{L}\left(\tau, z, \beta; \nu, \phi, a^{(\ell)}\right) : \left\|\left(a^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 \le 1, \left(a^{(\ell)}\right)_{g_{r,\ell}^c} = 0\right\}\right\}.$$

The dual function can then be written as

$$\begin{aligned}
g\left(\nu, \phi, a^{(\ell)}\right) &= \inf_{\tau,z,\beta}\mathcal{L}\left(\tau, z, \beta; \nu, \phi, a^{(\ell)}\right) \\
&= \inf_{\tau}\{-2\log\tau + \nu\tau\} + \inf_{z}\left\{\frac{1}{n}\|z\|_2^2 + \frac{1}{n}\langle\phi, z\rangle\right\} \\
&\quad + \inf_{\beta}\left\{-\nu\beta_r - \frac{1}{n}\left\langle\mathbf{X}_{1:r}^T\phi, \beta\right\rangle + \lambda\sum_{\ell=1}^{r-1}\left\langle W^{(\ell)} * a^{(\ell)}, \beta\right\rangle\right\} \\
&= 2\log\nu - 2\log 2 + 2 - \mathbb{1}_\infty\{\nu > 0\} - \frac{1}{4n}\|\phi\|_2^2 \\
&\quad - \mathbb{1}_\infty\left\{-\nu\mathbf{e}_r - \frac{1}{n}\mathbf{X}_{1:r}^T\phi + \lambda\sum_{\ell=1}^{r-1} W^{(\ell)} * a^{(\ell)} = 0\right\},
\end{aligned}$$

where $\mathbf{e}_r \in \mathbb{R}^r$ is such that $(\mathbf{e}_r)_r = 1$ and $(\mathbf{e}_r)_j = 0$ for all $j \neq r$. Thus the dual problem (up to a constant) is

$$\begin{aligned}
&\max_{\nu,\phi,a^{(\ell)}} g\left(\nu, \phi, a^{(\ell)}\right) \\
=& \min_{\nu,\phi,a^{(\ell)}}\left\{-2\log\nu + \frac{1}{4n}\|\phi\|_2^2 \quad \text{s.t.} \quad \nu > 0, \quad \left\|\left(a^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 \le 1, \left(a^{(\ell)}\right)_{g_{r,\ell}^c} = 0,\right. \\
&\qquad \left.\nu\mathbf{e}_r + \frac{1}{n}\mathbf{X}_{1:r}^T\phi = \lambda\sum_{\ell=1}^{r-1} W^{(\ell)} * a^{(\ell)}\right\}.
\end{aligned}$$

The primal-dual relation is

$$\hat{\beta}_r = \hat{\tau} = \frac{2}{\hat{\nu}} \qquad \hat{\phi} = -2\hat{z} = -2\mathbf{X}_{1:r}\hat{\beta}.$$

This implies that at optimal points

$$-\frac{2}{\hat{\beta}_r}\mathbf{e}_r + 2S_{1:r,1:r}\hat{\beta} + \lambda \sum_{\ell=1}^{r-1} W^{(\ell)} * \hat{a}^{(\ell)} = 0,$$

with $\left\|\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}}\right\|_2 \leq 1, \left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}^c} = 0.$

If we denote the objective function as

$$f(\beta) = -2\log\beta_r + \left\langle S_{1:r,1:r}, \beta\beta^T \right\rangle + \lambda P(\beta),$$

then from the equality $f(\hat{\beta}) = \mathcal{L}\left(\hat{\tau}, \hat{z}, \hat{\beta}; \hat{\nu}, \hat{\phi}, \hat{a}^{(\ell)}\right)$ together with the primal-dual relation, we have

$$P(\hat{\beta}) = \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * \hat{a}^{(\ell)}, \hat{\beta} \right\rangle = \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * \hat{\beta}, \hat{a}^{(\ell)} \right\rangle.$$

Suppose there exists some ℓ with $\hat{\beta}_{g_{r,\ell}} \neq 0$ but $\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}} \neq \frac{\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}}{\left\|\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}\right\|_2}$, then $\left\langle W^{(\ell)} * \hat{\beta}, \hat{a}^{(\ell)} \right\rangle < \left\|\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}\right\|_2$ while for other ℓ' by Cauchy-Schwarz inequality we have $\left\langle W^{(\ell')} * \hat{\beta}, \hat{a}^{(\ell')} \right\rangle \leq \left\|\left(W^{(\ell')} * \hat{\beta}\right)_{g_{r,\ell'}}\right\|_2$. Therefore, summing over all $\ell = 1, \ldots, r-1$ would give

$$P(\hat{\beta}) = \sum_{\ell=1}^{r-1} \left\|\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}\right\|_2 > \sum_{r=2}^{p} \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * \hat{\beta}, \hat{a}^{(\ell)} \right\rangle,$$

which leads to a contradiction. Thus $\left(\hat{a}^{(\ell)}\right)_{g_{r,\ell}} = \frac{\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}}{\left\|\left(W^{(\ell)} * \hat{\beta}\right)_{g_{r,\ell}}\right\|_2}$ for $\hat{\beta}_{g_{r,\ell}} \neq 0$ and $\left\|\hat{a}^{(\ell)}_{g_{r,\ell}}\right\|_2 \leq 1$ for $\hat{\beta}_{g_{r,\ell}} = 0$. □

A.11 Proof of Lemma 30

Proof. In this proof, we continue to use the notation in Appendix A.10. Observe that $\mathcal{L}\left(\tau, z, \beta; \nu, \phi, a^{(\ell)}\right)$ is jointly convex in τ, z and β, and it is strictly convex in τ and z. Thus, the minimizers $\hat{z}$ and $\hat{\tau}$ are unique.

To see this in a more general setting, without loss of generality, suppose $f(x, y)$ is convex in y and is strictly convex in x. Then for $x_1 \neq x_2$ and $\theta \in (0, 1)$ we have

$$f(\theta x_1 + (1-\theta) x_2, y) < \theta f(x_1, y) + (1-\theta) f(x_2, y)$$

Now suppose $(\hat{x}_1, \hat{y})$ and $(\hat{x}_2, \hat{y}_2)$ are both minima of f, then taking $\theta = 1/2$ we have $f\left(\frac{\hat{x}_1+\hat{x}_2}{2}, \hat{y}\right) < f(\hat{x}_1, \hat{y}) = f(\hat{x}_2, \hat{y})$, which leads to a contradiction.

By the primal-dual relation, we know that if $\hat{\beta}$ and $\tilde{\beta}$ are two solutions to (2.8), then $\hat{\beta}_r = \tilde{\beta}_r$ and $\mathbf{X}_{1:r}\hat{\beta} = \mathbf{X}_{1:r}\tilde{\beta}$. So from the equality $f(\hat{\beta}) = f(\tilde{\beta})$ we know that $P(\tilde{\beta}) = P(\hat{\beta})$. Also by

$$f\left(\hat{\beta}\right) = \mathcal{L}\left(\hat{\tau}, \hat{z}, \hat{\beta}; \hat{\nu}, \hat{\phi}, \hat{a}^{(\ell)}\right) \leq \mathcal{L}\left(\hat{\tau}, \hat{z}, \tilde{\beta}; \hat{\nu}, \hat{\phi}, \hat{a}^{(\ell)}\right) \leq \mathcal{L}\left(\tilde{\tau}, \tilde{z}, \tilde{\beta}; \tilde{\nu}, \tilde{\phi}, \tilde{a}^{(\ell)}\right) = f\left(\tilde{\beta}\right),$$

we have

$$\mathcal{L}\left(\hat{\tau}, \hat{z}, \hat{\beta}; \hat{\nu}, \hat{\phi}, \hat{a}^{(\ell)}\right) = \mathcal{L}\left(\hat{\tau}, \hat{z}, \tilde{\beta}; \hat{\nu}, \hat{\phi}, \hat{a}^{(\ell)}\right),$$

and thus

$$\sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * \hat{a}^{(\ell)}, \tilde{\beta} \right\rangle = \sum_{\ell=1}^{r-1} \left\langle W^{(\ell)} * \hat{a}^{(\ell)}, \hat{\beta} \right\rangle = P(\hat{\beta}) = P(\tilde{\beta}) = \sum_{\ell=1}^{r-1} \left\| \left(W^{(\ell)} * \tilde{\beta} \right)_{g_{r,\ell}} \right\|_2 .$$

Now for any $\ell \leq r-1$ suppose $\left\| \left(\hat{a}^{(\ell)} \right)_{g_{r,\ell}} \right\|_2 < 1$, then for the equality above to hold, we must have $\tilde{\beta}_{g_{r,\ell}} = 0$. Therefore, by Lemma 29, $\hat{\beta}_{g_{r,\ell}} = 0 \implies \tilde{\beta}_{g_{r,\ell}} = 0$, so any other solutions to (2.8) cannot be less sparse than $\hat{\beta}$. □

A.12 Proof of Lemma 31

Proof. By Lemma 30, any other solution β to (2.8) must have $\beta_{g_{J(\hat{\beta})}} = 0$. Recall that $J(\hat{\beta}) = r - 1 - K(\hat{\beta})$. The original problem (2.8) can thus be written equivalently

as

$$\min_{\gamma\in\mathbb{R}^{K(\hat{\beta})+1}} -2\log\gamma_{K(\hat{\beta})+1} + \frac{1}{n}\left\|\mathbf{X}_{\hat{S}}\gamma\right\|_2^2 + \lambda\sum_{\ell=1}^{K(\hat{\beta})}\left\|\left(\hat{W}^{(\ell)} * \gamma\right)_{g_{r,\ell}}\right\|_2,$$

where $\hat{W}^{(\ell)} = \left(W^{(\ell+\hat{J})}\right)_{\hat{S}}$.

Note that the penalty term is a convex function of γ. The Hessian matrix of the first term is a diagonal matrix of dimension $|\hat{\mathcal{S}}| = K(\hat{\beta}) + 1$ with non-negative entries in the diagonal. The Hessian matrix of the second term is $2S_{\hat{S}\hat{S}}$. Then by Assumption **A1**, the uniqueness follows from strict convexity. □

A.13 Proof of Lemma 33

Proof. Recall that

$$M_n = \frac{1}{n}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}^T \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} + \frac{4}{n^2\lambda^2}\tilde{\beta}_r^2\left\|\mathbf{O}_{\mathcal{I}}E_r\right\|_2^2.$$

We cite Lemma 9 (specifically in the form (60)) in Wainwright (2009) here for completeness.

Lemma 37 (Wainwright 2009)**.** *For $k \le n$, let $\mathbf{X}_{\mathcal{I}} \in \mathbb{R}^{n\times k}$ have i.i.d. rows from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ. If Σ has minimum eigenvalue $\kappa > 0$, then*

$$\mathrm{P}\left[\left\|\left(\frac{1}{n}X_{\mathcal{I}}^T X_{\mathcal{I}}\right)^{-1}\right\|_2 \ge \frac{9}{\kappa}\right] \le 2\exp\left(-\frac{n}{2}\right).$$

By the lemma above, Assumption **A4**, and (A.14)

$$\begin{aligned}\frac{1}{n}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}^T \left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} &\le \frac{9\kappa^2}{n}\left\|\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}\right\|^2 \\ &\le \frac{3\pi^2\kappa^2}{2}\frac{K}{n},\end{aligned}$$

with probability greater than $1 - 2\exp\left(-\frac{n}{2}\right)$.

Next we deal with the second term in M_n. Recall from (A.10) that

$$\begin{aligned}
\frac{4}{n^2\lambda^2}\tilde{\beta}_r^2 \left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2 &= \frac{4}{n^2}\left(\frac{\frac{1}{2}\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}} + \sqrt{\frac{1}{4}\left(\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right)^2 + \frac{4}{\lambda^2 n}\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2}}{\frac{2}{n}\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2}\right)^2 \left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2 \\
&\leq \frac{4}{n^2}\frac{\frac{1}{4}\left(\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right)^2 + \frac{4}{\lambda^2 n}\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2}{\frac{1}{n^2}\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^4}\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2 \\
&= \frac{\left(\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right)^2}{\left\|\mathbf{O}_{\mathcal{I}} E_r\right\|_2^2} + \frac{16}{\lambda^2 n}.
\end{aligned}$$

The next lemma gives us a handle on the numerator of the first term.

Lemma 38. *Using the general weight* (2.7)*, we have*

$$\mathrm{P}\left[\left|\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right| \geq 1\right] \leq 2\exp\left(-\frac{n\alpha^2}{3\theta\kappa^2\pi^2 K}\right) + 2\exp\left(-\frac{n}{2}\right).$$

Proof. Conditioned on $\mathbf{X}_{\mathcal{I}}$, from the decomposition (A.12) and the definition of $\mathbf{C}_{\mathcal{I}}$

$$\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}} = \Sigma_{r\mathcal{I}}\left(\Sigma_{\mathcal{I}\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} + E_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}.$$

By the irrepresentable assumption (**A3**) and (A.14),

$$\Sigma_{r\mathcal{I}}\left(\Sigma_{\mathcal{I}\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} \leq 1 - \alpha.$$

Note that $\mathrm{Var}\left(E_{ir}\right) = \theta_r^{(r)}$ for $i = 1, \ldots, n$. Let $B^{(r)} = E_r^T\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}$. By Lemma 37, $B^{(r)}$ has mean zero and variance at most

$$\mathrm{Var}\left(B^{(r)}\middle|\mathbf{X}_{\mathcal{I}}\right) = \frac{\theta_r^{(r)}}{n}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}}^T\left(\frac{1}{n}\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\left(\sum_{\ell=1}^{r-1} W^{(\ell)} * \tilde{a}^{(\ell)}\right)_{\mathcal{I}} \leq \frac{3\theta_r^{(r)}\kappa^2\pi^2 K}{2n},$$

with probability greater than $1 - 2\exp\left(\frac{n}{2}\right)$. By Lemma 32, we have that

$$\mathrm{P}\left[B^{(r)} \geq \alpha\right] \leq 2\exp\left(-\frac{n\alpha^2}{3\theta_r^{(r)}\kappa^2\pi^2 K}\right) + 2\exp\left(-\frac{n}{2}\right).$$

□

Since $\frac{\|\mathbf{O}_{\mathcal{I}}E_r\|_2^2}{\theta_r^{(r)}} \sim \chi^2(n-K)$. To bound it, we cite a concentration inequality from Wainwright (2009) (specifically (54b)) as the following lemma:

Lemma 39 (Tail Bounds for χ^2-variates, Wainwright 2009)**.** *For a centralized χ^2-variate X with d degrees of freedom, for all $\varepsilon \in (0, 1/2)$, we have*

$$\mathrm{P}\left[X \leq d(1-\varepsilon)\right] \leq \exp\left(-\frac{1}{4}d\varepsilon^2\right).$$

From Lemma 39 it follows that

$$\mathrm{P}\left[\|\mathbf{O}_{\mathcal{I}}E_r\|_2^2 \leq \theta_r^{(r)}(n-K)(1-\varepsilon)\right] \leq \exp\left(-\frac{1}{4}(n-K)\varepsilon^2\right),$$

which together with Lemma 38 implies that

$$\begin{aligned}
&\mathrm{P}\left[\frac{\left(\mathbf{X}_r^T\mathbf{C}_{\mathcal{I}}\right)^2}{\|\mathbf{O}_{\mathcal{I}}E_r\|_2^2} \geq \frac{1}{\theta_r^{(r)}(n-K)(1-\varepsilon)}\right] \\
\leq& 2\exp\left(-\frac{n\alpha^2}{3\theta_r^{(r)}\kappa^2\pi^2 K}\right) + 2\exp\left(-\frac{n}{2}\right) + \exp\left(-\frac{1}{4}(n-K)\varepsilon^2\right).
\end{aligned}$$

The result follows from a union bound. □

A.14 Proof of Lemma 34

Proof. The proof strategy is based on the proof of Lemma 2 in Bien et al. (2016).

For the design matrix $\mathbf{X}_{n\times p}$ with independent rows, denote $X_i = (\mathbf{X}_{i\cdot})^T \in \mathbb{R}^p$. Then X_i are i.i.d with mean 0 and true covariance matrix $\Sigma = \left(L^T L\right)^{-1}$ for $i = 1, ..., n$. And $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ has mean 0 and true covariance matrix $\frac{1}{n}\Sigma$.

Let $Y_i = LX_i \in \mathbb{R}^p$. Then Y_i are i.i.d with mean 0 and true covariance matrix $L\Sigma L^T = L\left(L^T L\right)^{-1} L^T = \mathbf{I}_p$. And $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n LX_i = L\bar{X}$ has mean zero and covariance matrix $\frac{1}{n}\mathbf{I}_p$. Also the corresponding design matrix $\mathbf{Y} = \mathbf{X}L^T$ has independent rows.

$$\begin{aligned} SL^T &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T L^T \\ &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(LX_i - L\bar{X}\right)^T = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)^T. \end{aligned}$$

So we have

$$\left(SL^T\right)_{ij} = n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \bar{X}_i\bar{Y}_j.$$

Letting

$$\mathcal{W} = SL^T - L^{-1},$$

we have that

$$\left|\mathcal{W}_{ij}\right| \le \left|n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \left(L^{-1}\right)_{ij}\right| + \left|\bar{X}_i\bar{Y}_j\right|.$$

$$\begin{aligned} &\mathrm{P}\left[\max_{ij}|\mathcal{W}|_{ij} > t\right] \\ \le &\mathrm{P}\left[\max_{ij}\left|n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \left(L^{-1}\right)_{ij}\right| > \frac{t}{2}\right] + \mathrm{P}\left[\max_{ij}\left|\bar{X}_i\bar{Y}_j\right| > \frac{t}{2}\right] \\ \le &\mathrm{P}\left[\left|n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \left(L^{-1}\right)_{ij}\right| > \frac{t}{2} \text{ for some } i,j\right] \\ &+ \mathrm{P}\left[\max_i\left|\bar{X}_i\right| > \sqrt{\frac{t}{2}}\right] + \mathrm{P}\left[\max_j\left|\bar{Y}_j\right| > \sqrt{\frac{t}{2}}\right] \\ \le &\sum_{ij}\mathrm{P}\left[\left|n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \left(L^{-1}\right)_{ij}\right| > \frac{t}{2}\right] + \sum_i \mathrm{P}\left[\left|\bar{X}_i\right| > \sqrt{\frac{t}{2}}\right] + \sum_j \mathrm{P}\left[\left|\bar{Y}_j\right| > \sqrt{\frac{t}{2}}\right] \\ \le &p^2\max_{ij}\mathrm{P}\left[\left|n^{-1}\sum_{k=1}^p X_{ki}Y_{kj} - \left(L^{-1}\right)_{ij}\right| > \frac{t}{2}\right] \\ &+ p\max_i \mathrm{P}\left[\left|\bar{X}_i\right| > \sqrt{\frac{t}{2}}\right] + p\max_j \mathrm{P}\left[\left|\bar{Y}_j\right| > \sqrt{\frac{t}{2}}\right] \\ := &p^2\max_{ij} I_{ij} + p\max_i I_i^X + p\max_j I_j^Y. \end{aligned}$$

Consider I_i^X first. Since X_{ki} are independent sub-Gaussian with variance Σ_{ii} for $k = 1, .., n$, we have

$$\begin{aligned}
\mathrm{E}\exp\left(t\frac{\bar{X}_i}{\sqrt{\Sigma_{ii}/n}}\right) &= \prod_{k=1}^{n}\mathrm{E}\exp\left(t\frac{X_{ki}}{\sqrt{n\Sigma_{ii}}}\right) \quad \text{by independence} \\
&\leq \prod_{k=1}^{n}\exp\left(\tilde{C}_1 t^2/n\right) = \exp(\tilde{C}_1 t^2) \quad \text{by the definition of sub-Gaussian,}
\end{aligned}$$

so $\bar{X}_i$ is sub-Gaussian with variance Σ_{ii}/n.

By Lemma 5.5 in Vershynin (2010), we have

$$\mathrm{P}\left[\left|\bar{X}_i\right| / \sqrt{\Sigma_{ii}^*} > t\right) \leq \exp\left(1 - t^2/K_1^2\right],$$

where K_1 is a constant that does not depend on i.

Following the same argument we have

$$\mathrm{E}\exp\left(t\bar{Y}_i/\sqrt{1/n}\right) = \prod_{k=1}^{n}\mathrm{E}\exp\left(tY_{ki}/\sqrt{n}\right) \leq \exp\left(\tilde{C}_2 t^2\right),$$

thus

$$\mathrm{P}\left[\left|\bar{Y}_i\right| / \sqrt{1/n} > t\right) \leq \exp\left(1 - t^2/K_2^2\right],$$

where K_2 is a constant that does not depend on i. And we have

$$\begin{aligned}
I_i^X + I_i^Y &= \mathrm{P}\left[\left|\bar{X}_i\right| > \sqrt{t/2}\right] + \mathrm{P}\left[\left|\bar{Y}_i\right| > \sqrt{t/2}\right] \\
&= \mathrm{P}\left[\frac{\left|\bar{X}_i\right|}{\sqrt{\Sigma_{ii}/n}} > \frac{\sqrt{t/2}}{\sqrt{\Sigma_{ii}/n}}\right] + \mathrm{P}\left[\left|\frac{\bar{Y}_i}{\sqrt{1/n}}\right| > \frac{\sqrt{t/2}}{\sqrt{1/n}}\right] \\
&\leq \exp\left(1 - \frac{nt}{2K_1^2\Sigma_{ii}^*}\right) + \exp\left(1 - \frac{nt}{2K_2^2}\right).
\end{aligned}$$

Thus

$$\max_i\left(I_i^X + I_i^Y\right) \leq 4\exp\left(-\frac{C_1 nt}{\max_i \Sigma_{ii}^*}\right) + 4\exp\left(-C_2 nt\right)$$

for some constant C_1.

Now consider the term I_{ij}. We have shown that both $\mathbf{X}$ and $\mathbf{Y}$ have independent rows. So for any i, j, $Z_k^{(ij)} = X_{ki}Y_{kj}$ are independent for $k = 1, \ldots, n$. Let $X \sim N(\mathbf{0}, \Sigma)$ and $Y \sim N\left(0, \mathbf{I}_p\right)$, then

$$\mathrm{E}\left(X_{ki}Y_{kj}\right) = \mathrm{Cov}\left(X, LX\right)_{ij} - 0 = \left[\mathrm{Cov}\left(X, X\right)L^T\right]_{ij} = \left(\Sigma L^T\right)_{ij} = \left(L^{-1}\right)_{ij}.$$

If there exist ν_{ij} and c_{ij} such that

$$\sum_{k=1}^{n} \mathrm{E}\left(X_{ki}^2 Y_{kj}^2\right) \leq \nu_{ij}$$
$$\sum_{k=1}^{n} \mathrm{E}\left\{\left(X_{ki}Y_{kj}\right)_+^q\right\} \leq \frac{q!}{2}\nu_{ij}c_{ij}^{q-2} \quad \text{for some} \quad q \geq 3 \in \mathbb{N},$$

then by Theorem 2.10 (Corollary 2.11) in Boucheron et al. (2013), $\forall t > 0$, we have

$$\mathrm{P}\left[\left|\sum_{k=1}^{n}\left(X_{ki}Y_{kj} - (L)_{ij}^{-1}\right)\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2\left(\nu_{ij} + c_{ij}t\right)}\right).$$

The rest of the proof focuses on characterizing ν_{ij} and c_{ij}. First, Lemma 5.5 in Vershynin (2010) shows that, for some constant K_3 that does not depend on j,

$$\left(\mathrm{E}\left|X_{ij}/\sqrt{\Sigma_{jj}}\right|^q\right)^{1/q} \leq K_3\sqrt{q}$$

holds for all $q \geq 1$. Thus,

$$\mathrm{E}\left|X_{ij}\right|^q \leq K_3^q q^{q/2}\left(\Sigma_{jj}\right)^{q/2}.$$

Following the same argument, there exists some constant K_4 that does not depend on j such that

$$\mathrm{E}\left|Y_{ij}\right|^q \leq K_4^q q^{q/2}$$

for all $q \geq 1$.

Therefore,

$$\sum_{k=1}^{n} \mathrm{E}\left(X_{ki}^2 Y_{kj}^2\right) \leq \sum_{k=1}^{n} \sqrt{\mathrm{E}X_{ki}^4 \mathrm{E}Y_{kj}^4} \leq n\sqrt{K_3^4 2^4 K_4^4 2^4 {\Sigma_{ii}}^2} = 16nK_3^2K_4^2\Sigma_{ii},$$

and

$$\sum_{k=1}^{n} \mathrm{E}\left\{\left(X_{ki}Y_{kj}\right)_{+}^{q}\right\} \leq \sum_{k=1}^{n} \sqrt{\mathrm{E}X_{ki}^{2q}\mathrm{E}Y_{kj}^{2q}} \leq n\sqrt{K_3^{2q}\,(2q)^{2q}\,K_4^{2q}\,(\Sigma_{ii})^2} = nK_3^qK_4^q\,(2q)^q\,(\Sigma_{ii})^{q/2}\,.$$

So taking

$$\nu_{ij} = K_5 n\Sigma_{ii}^*,$$
$$c_{ij} = K_5\sqrt{\Sigma_{ii}^*}$$

for some K_5 large enough and does not depend on i, j.

Now we have

$$I_{ij} \leq 2\exp\left(-\frac{n^2t^2}{4\left(2\nu_{ij} + c_{ij}tn\right)}\right) = 2\exp\left(-\frac{nt^2}{4\left(2K_5\Sigma_{ii}^* + K_5\sqrt{\Sigma_{ii}}t\right)}\right).$$

If $t \leq 2\max_i \sqrt{\Sigma_{ii}^*}$, then with $C_3 = (16K_5)^{-1}$ we have

$$I_{ij} \leq 2\exp\left(-\frac{C_2nt^2}{\max_i \Sigma_{ii}^*}\right).$$

To sum up, for any $0 < t \leq 2\max_i \sqrt{\Sigma_{ii}^*}$,

$$\mathrm{P}\left[\max_{ij}\left|\mathcal{W}_{ij}\right| > t\right] \leq 2p^2\exp\left(-\frac{C_2nt^2}{\max_i \Sigma_{ii}^*}\right) + 4p\exp\left(-\frac{C_1nt}{\max_i \Sigma_{ii}^*}\right) + 4p\exp\left(-C_2nt\right).$$

□

APPENDIX B

APPENDIX OF CHAPTER 3

B.1 Proof of Lemma 8

From (3.2) in the paper, it follows that

$$\hat{\sigma}_\lambda^2 \leq \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta^*\|_2^2 + 2\lambda \|\beta^*\|_1 = \frac{1}{n} \|\varepsilon\|_2^2 + 2\lambda \|\beta^*\|_1 .$$

By introducing the dual variable $2u/n \in \mathbb{R}^n$,

$$\begin{aligned}
\hat{\sigma}_\lambda^2 &= \min_\beta \left(\frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + 2\lambda \|\beta\|_1 \right) = \min_{\beta,\mathbf{z}} \max_u \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{z}\|_2^2 + \frac{2}{n} u^T (\mathbf{z} - \mathbf{X}\beta) + 2\lambda \|\beta\|_1 \right\} \\
&\geq \max_u \min_{\beta,\mathbf{z}} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{z}\|_2^2 + \frac{2}{n} u^T (\mathbf{z} - \mathbf{X}\beta) + 2\lambda \|\beta\|_1 \right\} \\
&= \max_u \left(\frac{1}{n} \|\mathbf{y}\|_2^2 - \frac{1}{n} \|\mathbf{y} - u\|_2^2 , \text{subject to } \left\| \mathbf{X}^T u \right\|_\infty \leq n\lambda \right).
\end{aligned}$$

By assumption, ε is dual feasible, which means that

$$\hat{\sigma}_\lambda^2 \geq \frac{1}{n} \|\mathbf{y}\|_2^2 - \frac{1}{n} \|\mathbf{y} - \varepsilon\|_2^2 \geq \frac{1}{n} \|\varepsilon\|_2^2 + \frac{2}{n} \varepsilon^T \mathbf{X}\beta^* \geq \frac{1}{n} \|\varepsilon\|_2^2 - 2\lambda \|\beta^*\|_1 ,$$

where in the last step we applied Hölder's inequality.

B.2 Proof of Propositions 7 and 14

We prove in this section that both the natural lasso and the organic lasso estimates of error variance can be simply expressed as the minimizing values of certain convex optimization problems. To do so, we exploit the first order optimality condition of each convex program.

We start with proving that the natural lasso estimate of σ^2 is the minimal value of a lasso problem (3.2). The following lemma characterizes the conditions for which $(\hat{\theta}_\lambda, \hat{\phi}_\lambda)$ is a solution to (3.8) with $\Omega(\theta, \phi) = \|\theta\|_1$.

Lemma 40 (Optimality condition of the natural lasso)**.** *For any $\lambda > 0$, $(\hat{\theta}_\lambda, \hat{\phi}_\lambda)$ is a solution to* (3.8) *with $\Omega(\theta, \phi) = \|\theta\|_1$ if and only if*

$$-\frac{1}{\hat{\phi}_\lambda} + \frac{1}{n}\|\mathbf{y}\|_2^2 - \frac{\left\|\mathbf{X}\hat{\theta}_\lambda\right\|_2^2}{n\hat{\phi}_\lambda^2} = 0, \qquad -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\frac{\hat{\theta}_\lambda}{\hat{\phi}_\lambda} + n\lambda\hat{g} = \mathbf{0}$$

where $\hat{g} \in \partial(\|\hat{\theta}_\lambda\|_1)$.

Given $(\hat{\theta}_\lambda, \hat{\phi}_\lambda)$, we reverse the natural parameterization to get $\hat{\beta}_\lambda = \hat{\phi}_\lambda^{-1}\hat{\theta}_\lambda$ and $\hat{\sigma}_\lambda^2 = \hat{\phi}_\lambda^{-1}$. From Lemma 40,

$$\hat{\sigma}_\lambda^2 = \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2\right) \quad \text{and} \quad 0 = -\hat{\beta}_\lambda^T\mathbf{X}^T\mathbf{y} + \left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2 + n\lambda\left\|\hat{\beta}_\lambda\right\|_1.$$

Note that

$$\left\|\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\right\|_2^2 = \|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2 + 2\left(\left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2 - \mathbf{y}^T\mathbf{X}\hat{\beta}_\lambda\right) = \|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2 - 2n\lambda\left\|\hat{\beta}_\lambda\right\|_1.$$

We have

$$\hat{\sigma}_\lambda^2 = \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\hat{\beta}_\lambda\right\|_2^2\right) = \frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\right\|_2^2 + 2\lambda\left\|\hat{\beta}_\lambda\right\|_1.$$

We show that the organic lasso estimate of σ^2 is the minimal value of the ℓ_1^2-penalized least squares problem. As the natural lasso, we start with studying the following optimality condition:

Lemma 41 (Optimality condition of the organic lasso)**.** *For any $\lambda > 0$, $\left(\check{\theta}_\lambda, \check{\phi}_\lambda\right)$ is a solution to* (3.15) *if and only if*

$$-\frac{1}{\check{\phi}_\lambda} + \frac{1}{n}\|\mathbf{y}\|_2^2 - \frac{\left\|\mathbf{X}\check{\theta}_\lambda\right\|_2^2}{n\check{\phi}_\lambda^2} - 2\lambda\frac{\left\|\check{\theta}_\lambda\right\|_1^2}{\check{\phi}_\lambda^2} = 0, \qquad -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\frac{\check{\theta}_\lambda}{\check{\phi}_\lambda} + 2n\lambda\frac{\left\|\check{\theta}_\lambda\right\|_1}{\check{\phi}_\lambda}\check{g} = \mathbf{0}$$

where $\check{g} \in \partial(\|\check{\theta}\|_1)$.

So following the natural parameterization, we have that $\check{\beta}_\lambda = \check{\theta}_\lambda^{-1}\check{\rho}_\lambda$ and $\check{\sigma}_\lambda^2 = \check{\rho}_\lambda^{-1}$, and

$$\begin{aligned}\check{\sigma}_\lambda^2 &= \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 - 2n\lambda\left\|\check{\beta}_\lambda\right\|_1^2\right)\\ 0 &= -\check{\beta}_\lambda^T\mathbf{X}^T\mathbf{y} + \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 + 2n\lambda\left\|\check{\beta}_\lambda\right\|_1^2.\end{aligned}$$

Note that

$$\begin{aligned}\left\|\mathbf{y} - \mathbf{X}\check{\beta}_\lambda\right\|_2^2 &= \|\mathbf{y}\|_2^2 + \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 - 2\mathbf{y}^T\mathbf{X}\check{\beta}_\lambda\\ &= \|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 + 2\left(\left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 - \mathbf{y}^T\mathbf{X}\check{\beta}_\lambda\right)\\ &= \|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 - 4n\lambda\left\|\check{\beta}_\lambda\right\|_1^2.\end{aligned}$$

We have

$$\check{\sigma}_\lambda^2 = \frac{1}{n}\left(\|\mathbf{y}\|_2^2 - \left\|\mathbf{X}\check{\beta}_\lambda\right\|_2^2 - 2n\lambda\left\|\check{\beta}_\lambda\right\|_1^2\right) = \frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\check{\beta}_\lambda\right\|_2^2 + 2\lambda\left\|\check{\beta}_\lambda\right\|_1^2.$$

B.3 Proof of Lemma 16: the dual problem of the ℓ_1^2-penalized least squares

The primal problem of the ℓ_1^2-penalized least squares (3.16) in the paper can be written as an equality constrained minimization problem:

$$\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{n}\|\mathbf{y} - \mathbf{z}\|_2^2 + 2\lambda\|\beta\|_1^2 \quad \text{s.t.} \quad \frac{2}{n}\mathbf{z} = \frac{2}{n}\mathbf{X}\beta\right).$$

The Lagrange dual function is

$$\begin{aligned}g(u) &= \min_{\beta\in\mathbb{R}^p,\mathbf{z}\in\mathbb{R}^n}\left\{\frac{1}{n}\|\mathbf{y} - \mathbf{z}\|_2^2 + 2\lambda\|\beta\|_1^2 + \frac{2u^T}{n}(\mathbf{z} - \mathbf{X}\beta)\right\}\\ &= \min_{\mathbf{z}\in\mathbb{R}^n}\left(\frac{1}{n}\|\mathbf{y} - \mathbf{z}\|_2^2 + \frac{2}{n}u^T\mathbf{z}\right) + \min_{\beta\in\mathbb{R}^p}\left\{2\lambda\|\beta\|_1^2 - 2\left(\frac{X^Tu}{n}\right)^T\beta\right\}.\end{aligned}$$

The minimization of u is

$$\min_{\mathbf{z}\in\mathbb{R}^n}\left(\frac{1}{n}\left\|\mathbf{y}-\mathbf{z}\right\|_2^2+\frac{2}{n}u^T\mathbf{z}\right)=\frac{2}{n}u^T\mathbf{y}-\frac{1}{n}\left\|u\right\|_2^2=\frac{1}{n}\left(\left\|\mathbf{y}\right\|_2^2-\left\|\mathbf{y}-u\right\|_2^2\right),$$

where the minimum is attained at

$$\hat{\mathbf{z}}=\mathbf{y}-u.$$

The minimization problem of β can be written as

$$\min_{\beta\in\mathbb{R}^p}\left\{2\lambda\left\|\beta\right\|_1^2-2\left(\frac{\mathbf{X}^Tu}{n}\right)^T\beta\right\}=-2\lambda\max_{\beta\in\mathbb{R}^p}\left\{\left(\frac{\mathbf{X}^Tu}{\lambda n}\right)^T\beta-\left\|\beta\right\|_1^2\right\}.$$

Observe that the maximum is the Fenchel conjugate function of $\left\|\cdot\right\|_1^2$, evaluated at $(\lambda n)^{-1}\mathbf{X}^Tu$. By Boyd & Vandenberghe (2004, Example 3.27, pp. 92-93),

$$-2\lambda\max_{\beta\in\mathbb{R}^p}\left\{\left(\frac{\mathbf{X}^Tu}{\lambda n}\right)^T\beta-\left\|\beta\right\|_1^2\right\}=-\frac{2\lambda}{4}\left\|\frac{\mathbf{X}^Tu}{\lambda n}\right\|_\infty^2=-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^Tu}{n}\right\|_\infty^2.$$

So

$$g\left(u\right)=\frac{1}{n}\left(\left\|\mathbf{y}\right\|_2^2-\left\|\mathbf{y}-u\right\|_2^2\right)-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^Tu}{n}\right\|_\infty^2.$$

B.4 Proof of Lemma 17

A direct upper bound is

$$\check{\sigma}_\lambda^2\leq\frac{1}{n}\left\|\mathbf{y}-\mathbf{X}\beta^*\right\|_2^2+2\lambda\left\|\beta^*\right\|_1^2=\frac{1}{n}\left\|\varepsilon\right\|_2^2+2\lambda\left\|\beta^*\right\|_1^2.$$

To get a lower bound of $\hat{\sigma}^2$, note that the dual problem in Lemma 16 and the strong duality imply that

$$\begin{aligned}\check{\sigma}_\lambda^2&=\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{n}\left\|\mathbf{y}-\mathbf{X}\beta\right\|_2^2+2\lambda\left\|\beta\right\|_1^2\right)=\max_{u\in\mathbb{R}^n}\left(\frac{1}{n}\left\|\mathbf{y}\right\|_2^2-\frac{1}{n}\left\|\mathbf{y}-u\right\|_2^2-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^Tu}{n}\right\|_\infty^2\right)\\&\geq\frac{1}{n}\left\|\mathbf{y}\right\|_2^2-\frac{1}{n}\left\|\mathbf{y}-\varepsilon\right\|_2^2-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^T\varepsilon}{n}\right\|_\infty^2=\frac{1}{n}\left\|\varepsilon\right\|_2^2+\frac{2}{n}\varepsilon^T\mathbf{X}\beta^*-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^T\varepsilon}{n}\right\|_\infty^2\\&\geq\frac{1}{n}\left\|\varepsilon\right\|_2^2-2\left\|\frac{\mathbf{X}^T\varepsilon}{n}\right\|_\infty\left\|\beta^*\right\|_1-\frac{1}{2\lambda}\left\|\frac{\mathbf{X}^T\varepsilon}{n}\right\|_\infty^2\geq\frac{1}{n}\left\|\varepsilon\right\|_2^2-2\lambda\sigma^2\left(\frac{\left\|\beta^*\right\|_1}{\sigma}+\frac{1}{4}\right),\end{aligned}$$

where the last inequality holds for

$$\lambda \geq \frac{\left\|\mathbf{X}^T \varepsilon\right\|_\infty}{n\sigma}.$$

B.5 Proof of Theorem 9 and Theorem 18

We present in this section the proof of Theorem 18. The proof of Theorem 9 follows the same set of arguments. First we use the following lemma to characterize the event that $\lambda \geq n^{-1}\sigma^{-1}\|\mathbf{X}^T\varepsilon\|_\infty$ is true, so that we can use Lemma 17 to prove a high probability bound.

Lemma 42 (Corollary 4.3, Giraud (2014))**.** *Assume that each column $\mathbf{X}_j$ of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfies $\|\mathbf{X}_j\|_2^2 = n$ for all $j = 1, \ldots, p$, and $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$. Then for any $L > 0$,*

$$\mathrm{P}\left\{\frac{\left\|\mathbf{X}^T\varepsilon\right\|_\infty}{n\sigma} > \left(\frac{2\log p + 2L}{n}\right)^{1/2}\right\} \leq e^{-L}.$$

Lemma 42 implies that a good choice of the value of λ would be $\{n^{-1}(2\log p + 2L)\}^{1/2}$, which does not depend on any parameter of the underlying model. The following corollary shows that with this value of λ, the organic lasso estimate of σ^2 is close to the oracle estimator with high probability.

Corollary 43. *Assume that each column $\mathbf{X}_j$ of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfies $\|\mathbf{X}_j\|_2^2 = n$ for all $j = 1, \ldots, p$, and $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$. Then for any $L > 0$, the organic lasso with*

$$\lambda = \left(\frac{2\log p + 2L}{n}\right)^{1/2}$$

has the following bound

$$\left(\breve{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2 \leq 8\max\left\{\|\beta^*\|_1^2, \sigma^2\left(\frac{\|\beta^*\|_1}{\sigma} + \frac{1}{4}\right)\right\}^2 \frac{\log p + L}{n}$$

with probability greater than $1 - e^{-L}$.

In general, a high probability bound does not necessarily imply an expectation bound. However, when the probability bound holds with an exponential tail, it implies an expectation bound with essentially the same rate.

Theorem 44. *Assume that each column* $\mathbf{X}_j$ *of the design matrix* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *satisfies* $\left\|\mathbf{X}_j\right\|_2^2 = n$ *for all* $j = 1, \ldots, p$, *and* $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$. *Then, for any constant* $M > 1$, *the organic lasso estimate with*

$$\lambda = \left(\frac{2M\log p}{n}\right)^{1/2}$$

satisfies the following bound in expectation:

$$\mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\left\|\varepsilon\right\|_2^2\right)^2\right\} \leq 8\left(M + \frac{p^{1-M}}{\log p}\right)\max\left\{\left\|\beta^*\right\|_1^2, \sigma^2\left(\frac{\left\|\beta^*\right\|_1}{\sigma} + \frac{1}{4}\right)\right\}^2 \frac{\log p}{n}.$$

Proof. For any $M > 1$, take $L = (M-1)\log p$ in Corollary 43. Denote $X_n = (\check{\sigma}_\lambda^2 - n^{-1}\|\varepsilon\|^2)^2$, and $r_n = 8\max(\|\beta^*\|_1^2, \sigma\|\beta^*\|_1 + \sigma^2/4)^2 n^{-1}\log p$. Then we have

$$\mathrm{P}\left(X_n > Mr_n\right) \leq e^{-(M-1)\log p}.$$

So

$$\begin{aligned}\mathrm{E}\left(\frac{X_n}{r_n}\right) &= \int_0^\infty \mathrm{P}\left(\frac{X_n}{r_n} > t\right)\mathrm{d}t = \int_0^M \mathrm{P}\left(\frac{X_n}{r_n} > t\right)\mathrm{d}t + \int_M^\infty \mathrm{P}\left(\frac{X_n}{r_n} > t\right)\mathrm{d}t \\ &\leq M + \int_M^\infty e^{-(t-1)\log p}\mathrm{d}t = M + \frac{p^{1-M}}{\log p},\end{aligned}$$

and the expectation bound follows. □

Now we are ready to present the proof of Theorem 18. Since $\|\varepsilon\|_2^2/\sigma^2 \sim \chi^2(n)$, we have

$$\mathrm{E}\left(\frac{1}{n}\left\|\varepsilon\right\|_2^2\right) = \sigma^2, \qquad \mathrm{Var}\left(\frac{1}{n}\left\|\varepsilon\right\|_2^2\right) = \frac{2\sigma^4}{n},$$

Therefore,

$$
\begin{aligned}
\mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \sigma^2\right)^2\right\} &= \mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2 + \frac{1}{n}\|\varepsilon\|_2^2 - \sigma^2\right)^2\right\} \\
&= \mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2\right\} + \mathrm{E}\left\{\left(\frac{1}{n}\|\varepsilon\|_2^2 - \sigma^2\right)^2\right\} + 2\mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)\left(\frac{1}{n}\|\varepsilon\|_2^2 - \sigma^2\right)\right\} \\
&\leq \mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2\right\} + \mathrm{Var}\left(\frac{1}{n}\|\varepsilon\|_2^2\right) + 2\left\{\mathrm{Var}\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)\mathrm{Var}\left(\frac{1}{n}\|\varepsilon\|_2^2\right)\right\}^{1/2} \\
&\leq \mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2\right\} + \mathrm{Var}\left(\frac{1}{n}\|\varepsilon\|_2^2\right) + 2\left[\mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2\right\}\mathrm{Var}\left(\frac{1}{n}\|\varepsilon\|_2^2\right)\right]^{1/2} \\
&= \left[\left[\mathrm{E}\left\{\left(\check{\sigma}_\lambda^2 - \frac{1}{n}\|\varepsilon\|_2^2\right)^2\right\}\right]^{1/2} + \left\{\mathrm{Var}\left(\frac{1}{n}\|\varepsilon\|_2^2\right)\right\}^{1/2}\right]^2 \\
&\leq \left[\left\{8\left(M + \frac{p^{1-M}}{\log p}\right)\right\}^{1/2} \max\left\{\|\beta^*\|_1^2, \sigma^2\left(\frac{\|\beta^*\|_1}{\sigma} + \frac{1}{4}\right)\right\}\left(\frac{\log p}{n}\right)^{1/2} + \sigma^2\left(\frac{2}{n}\right)^{1/2}\right]^2,
\end{aligned}
$$

where the last inequality holds from Theorem 44.

B.6 Proof of Remark 11

For the independent zero-mean noise ε_i with variance σ^2 and bounded m-th order moment ($m = 3, 4, \dots$)

$$
\mathrm{E}|\varepsilon_i|^m \leq \frac{m!}{2}K^{m-2}
$$

for some constant $K > 0$, a Bernstein's type inequality (Bühlmann & Van De Geer 2011, Lemma 14.13) implies that

$$
\mathrm{P}\left[\max_{1\leq j\leq p}\frac{1}{n\sigma}\left\|\mathbf{X}_j^T\varepsilon\right\|_\infty \geq \frac{2K\log p}{n} + 2\left\{\frac{\log(2p)}{n}\right\}^{1/2}\right] \leq \frac{1}{p}.
$$

Then the proof of Corollary 43 goes through.

B.7 Proof of Proposition 12 and Proposition 13

The following lemma gives a general result on the estimation error of $\hat{\sigma}^2$ of the form (3.3) in the paper based on $\hat{\beta}$:

Lemma 45.

$$\left|\hat{\sigma}^2 - \frac{1}{n}\|\varepsilon\|_2^2\right| \leq \frac{1}{n}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\right\|_2^2 + \frac{2}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left(\|\beta^*\|_1 + \|\hat{\beta}\|_1\right)$$

Proof. First by definition

$$\hat{\sigma}^2 = \frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|_2^2 = \frac{1}{n}\left\|\varepsilon + \mathbf{X}\beta^* - \mathbf{X}\hat{\beta}\right\|_2^2 = \frac{1}{n}\|\varepsilon\|_2^2 + \frac{1}{n}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\right\|_2^2 + \frac{2}{n}\varepsilon^T\mathbf{X}\left(\hat{\beta} - \beta^*\right).$$

Note that

$$\left|\varepsilon^T\mathbf{X}\left(\hat{\beta} - \beta^*\right)\right| \leq \|\mathbf{X}^T\varepsilon\|_\infty\|\hat{\beta} - \beta^*\|_1,$$

and the result follows. □

B.7.1 Slow rate bound for the naive estimator of σ^2

We now give the proof of Proposition 12. From the basic inequality

$$\frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\right\|_2^2 + 2\lambda\|\hat{\beta}_\lambda\|_1 \leq \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta^*\|_2^2 + 2\lambda\|\beta^*\|_1,$$

which implies that

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\|\hat{\beta}_\lambda\|_1 &\leq \frac{2}{n}\left|\varepsilon^T\mathbf{X}\left(\hat{\beta}_\lambda - \beta^*\right)\right| + 2\lambda\|\beta^*\|_1 \\ &\leq \frac{2}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left\|\hat{\beta}_\lambda - \beta^*\right\|_1 + 2\lambda\|\beta^*\|_1.\end{aligned}$$

We thank Irina Gaynanova (Gaynanova n.d.) for showing us the technique of taking λ to be twice its usual size. For $\lambda \geq 2n^{-1}\|\mathbf{X}^T\varepsilon\|_\infty$, we have that

$$\frac{1}{n}\left\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\|\hat{\beta}_\lambda\|_1 \leq \lambda\|\hat{\beta}_\lambda - \beta^*\|_1 + 2\lambda\|\beta^*\|_1 \leq \lambda\|\hat{\beta}_\lambda\|_1 + 3\lambda\|\beta^*\|_1,$$

so $n^{-1}\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta^*\|_2^2 + \lambda\|\hat{\beta}_\lambda\|_1 \leq 3\lambda\|\beta^*\|_1$. So by Lemma 45 we have

$$\begin{aligned}\left|\hat{\sigma}^2_{\text{naive}} - \frac{1}{n}\|\varepsilon\|_2^2\right| &\leq \frac{1}{n}\left\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 + \frac{2}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left(\|\beta^*\|_1 + \|\hat{\beta}_\lambda\|_1\right)\\ &\leq \frac{1}{n}\left\|\mathbf{X}\hat{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 + \lambda\|\beta^*\|_1 + \lambda\|\hat{\beta}_\lambda\|_1 \leq 4\lambda\|\beta^*\|_1.\end{aligned}$$

Finally, taking $\lambda = 2\sigma\{n^{-1}(2\log p + 2L)\}^{1/2}$ with $L = \log p$, the result follows from Lemma 42.

B.7.2 Slow rate bound for the square-root/scaled lasso estimator of σ^2

As shown in Lederer et al. (2016) (proof of Lemma A.3), we note that with probability 1, $\|y - X\tilde{\beta}_{\text{SQRT}}\|_2 > 0$ for $\lambda > 0$. So the first order optimality condition of the square-root/scaled lasso is

$$\frac{1}{\sqrt{n}}\frac{-\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right)}{\left\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right\|_2} + \lambda\hat{g} = 0$$

for some $\hat{g} \in \partial\|\tilde{\beta}_{\text{SQRT}}\|_1$. Taking an inner product with $\tilde{\beta}_{\text{SQRT}} - \beta^*$ on both sides, we have

$$-\frac{1}{\sqrt{n}}\frac{\left(\tilde{\beta}_{\text{SQRT}} - \beta^*\right)^T\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right)}{\left\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right\|_2} + \lambda\hat{g}^T\left(\tilde{\beta}_{\text{SQRT}} - \beta^*\right) = 0,$$

which implies that

$$\frac{\left\|\mathbf{X}\left(\beta^* - \tilde{\beta}_{\text{SQRT}}\right)\right\|_2^2}{\sqrt{n}\left\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right\|_2} - \frac{\left(\tilde{\beta}_{\text{SQRT}} - \beta^*\right)^T\mathbf{X}^T\varepsilon}{\sqrt{n}\left\|\mathbf{y} - \mathbf{X}\tilde{\beta}_{\text{SQRT}}\right\|_2} \leq \lambda\hat{g}^T\left(\beta^* - \tilde{\beta}_{\text{SQRT}}\right) \leq \lambda\|\beta^*\|_1 - \lambda\|\tilde{\beta}_{\text{SQRT}}\|_1,$$

and thus

$$\begin{aligned}
&\frac{1}{n}\left\|\mathbf{X}\left(\beta^*-\tilde{\beta}_{\mathrm{SQRT}}\right)\right\|_2^2 \leq \frac{1}{n}\left|\varepsilon^T\mathbf{X}\left(\tilde{\beta}_{\mathrm{SQRT}}-\beta^*\right)\right| + \frac{\lambda}{\sqrt{n}}\left\|\mathbf{y}-\mathbf{X}\tilde{\beta}_{\mathrm{SQRT}}\right\|_2\left(\|\beta^*\|_1-\|\tilde{\beta}_{\mathrm{SQRT}}\|_1\right)\\
\leq&\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left\|\tilde{\beta}_{\mathrm{SQRT}}-\beta^*\right\|_1 + \frac{\lambda}{\sqrt{n}}\left\|\mathbf{y}-\mathbf{X}\tilde{\beta}_{\mathrm{SQRT}}\right\|_2\left(\|\beta^*\|_1-\|\tilde{\beta}_{\mathrm{SQRT}}\|_1\right)\\
\leq&\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left(\|\tilde{\beta}_{\mathrm{SQRT}}\|_1+\|\beta^*\|_1\right) + \frac{\lambda}{\sqrt{n}}\left\|\mathbf{y}-\mathbf{X}\tilde{\beta}_{\mathrm{SQRT}}\right\|_2\left(\|\beta^*\|_1-\|\tilde{\beta}_{\mathrm{SQRT}}\|_1\right)\\
\leq&\left(\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty+\frac{\lambda}{\sqrt{n}}\left\|\mathbf{y}-\mathbf{X}\tilde{\beta}_{\mathrm{SQRT}}\right\|_2\right)\|\beta^*\|_1+\left(\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty-\frac{\lambda}{\sqrt{n}}\left\|y-X\tilde{\beta}_{\mathrm{SQRT}}\right\|_2\right)\|\tilde{\beta}_{\mathrm{SQRT}}\|_1.
\end{aligned}$$

Taking $\lambda = 3n^{-1/2}\|y - X\tilde{\beta}_{\mathrm{SQRT}}\|_2^{-1}\|X^T\varepsilon\|_\infty$, which is 3 times what is suggested in Lederer et al. (2016), we have

$$\frac{1}{n}\left\|\mathbf{X}\left(\beta^*-\tilde{\beta}_{\mathrm{SQRT}}\right)\right\|_2^2 \leq \frac{4\left\|\mathbf{X}^T\varepsilon\right\|_\infty}{n}\|\beta^*\|_1 - \frac{2\|X^T\varepsilon\|_\infty}{n}\|\tilde{\beta}_{\mathrm{SQRT}}\|_1.$$

By Lemma 45

$$\begin{aligned}
\left|\tilde{\sigma}^2_{\mathrm{SQRT}}-\frac{1}{n}\|\varepsilon\|_2^2\right| &\leq \frac{1}{n}\left\|\mathbf{X}\tilde{\beta}_{\mathrm{SQRT}}-\mathbf{X}\beta^*\right\|_2^2+\frac{2}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\left(\|\beta^*\|_1+\|\tilde{\beta}_{\mathrm{SQRT}}\|_1\right)\\
&\leq \frac{6}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty\|\beta^*\|_1.
\end{aligned}$$

The result then follows from Lemma 42 by taking $L = \log p$.

B.8 Proof of Proposition 15: scale-equivariance of the organic lasso

Proof. Suppose $\check{\beta}_\lambda(\mathbf{y})$ is a solution to the organic lasso, where we write out explicitly the dependence of the solution on the response $\mathbf{y}$. Then using notation from previous section,

$$\begin{aligned}
L\left(t\check{\beta}_\lambda(\mathbf{y})\,|t\mathbf{y},\lambda\right) &= \frac{1}{n}\left\|t\mathbf{y}-t\mathbf{X}\check{\beta}_\lambda(\mathbf{y})\right\|_2^2+2\lambda\left\|t\check{\beta}_\lambda(\mathbf{y})\right\|_1^2\\
&= t^2L\left(\check{\beta}_\lambda(\mathbf{y})\,|\mathbf{y},\lambda\right).
\end{aligned}$$

This implies that $t\check{\beta}_\lambda(\mathbf{y})$ is a solution to the problem with response $t\mathbf{y}$, i.e., $\check{\beta}_\lambda(t\mathbf{y}) = t\check{\beta}(\mathbf{y})$. Consequently,

$$\begin{aligned}\check{\sigma}^2_\lambda(t\mathbf{y}) &= \min_\beta L(\beta_\lambda | t\mathbf{y}, \lambda) \\ &= L\left(t\check{\beta}_\lambda(\mathbf{y}, \lambda) | t\mathbf{y}, \lambda\right) = t^2 L\left(\check{\beta}_\lambda(\mathbf{y}, \lambda) | \mathbf{y}, \lambda\right) = t^2 \check{\sigma}^2_\lambda(\mathbf{y}, \lambda),\end{aligned}$$

which establishes the theorem. □

B.9 Proof of Theorem 19

Proof. We start from the basic inequality

$$\frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\check{\beta}_\lambda\right\|_2^2 + 2\lambda\left\|\check{\beta}_\lambda\right\|_1^2 \leq \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta^*\|_2^2 + 2\lambda\|\beta^*\|_1^2,$$

which leads to

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\check{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 &\leq 2\left(\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right)^T\left(\check{\beta}_\lambda - \beta^*\right) + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right) \\ &\leq 2\left\|\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right\|_\infty\left\|\check{\beta}_\lambda - \beta^*\right\|_1 + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right).\end{aligned}$$

If

$$\left\|\frac{\mathbf{X}^T\boldsymbol{\varepsilon}}{n}\right\|_\infty \leq \sigma\lambda,$$

then

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\check{\beta}_\lambda - \mathbf{X}\beta^*\right\|_2^2 &\leq 2\sigma\lambda\left\|\check{\beta}_\lambda - \beta^*\right\|_1 + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right) \\ &\leq \sigma^2\lambda + \lambda\left\|\check{\beta}_\lambda - \beta^*\right\|_1^2 + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right) \\ &\leq \sigma^2\lambda + \lambda\left(\left\|\check{\beta}_\lambda\right\|_1 + \|\beta^*\|_1\right)^2 + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right) \\ &\leq \sigma^2\lambda + 2\lambda\left(\left\|\check{\beta}_\lambda\right\|_1^2 + \|\beta^*\|_1^2\right) + 2\lambda\left(\|\beta^*\|_1^2 - \left\|\check{\beta}_\lambda\right\|_1^2\right) \\ &= \sigma^2\lambda + 4\lambda\|\beta^*\|_1^2.\end{aligned}$$

The result then holds from Lemma 42. □

B.10 Mapping between the paths of the natural and organic lasso

In this section, we draw a connection between the natural lasso and the organic lasso estimates of β^*.

Theorem 46. *Letting $\hat{\beta}_s$ and $\check{\beta}_t$ denote the lasso and organic lasso estimates of β^* with tuning parameters s and t,*

$$\hat{\beta}_\lambda = \check{\beta}_{\lambda/(2\|\hat{\beta}_\lambda\|_1)}, \qquad \check{\beta}_\nu = \hat{\beta}_{2\nu\|\check{\beta}_\nu\|_1}. \tag{B.1}$$

This result implies that one can start with a lasso solution $\hat{\beta}_\lambda$ with tuning parameter λ, and then report a solution to the organic lasso with tuning parameter $(2\|\hat{\beta}_\lambda\|_1)^{-1}\lambda$. Likewise, an organic lasso solution $\check{\beta}_\nu$ is equivalent to a standard lasso solution with tuning parameter $2\nu\|\check{\beta}_\nu\|_1$. This equivalence is also observed in Lorbert et al. (2010) that considers a more general penalty.

Although the methods' paths are the same, this does not imply that the cross-validated methods will be the same. In K-fold cross-validation, the natural lasso estimator is evaluated on K differing datasets for a fixed value of λ. A fixed tuning parameter λ for the natural lasso over multiple datasets corresponds to running the organic lasso with a different λ on each fold. Thus, the two methods in fact have different cross-validation performance.

Proof. Let $\hat{\beta}_\lambda$ be a solution to (3.2) with tuning parameter λ, and $\tilde{\beta}_\nu$ be a solution to (3.16) with tuning parameter ν, then they satisfy optimality conditions

$$-\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda\right) + \lambda\hat{g} = \mathbf{0} \qquad \text{where} \qquad \hat{g} \in \partial\left(\left\|\hat{\beta}_\lambda\right\|_1\right), \tag{B.2}$$

$$-\frac{1}{n}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\tilde{\beta}_\nu\right) + 2\nu\left\|\tilde{\beta}_\nu\right\|_1\tilde{g} = \mathbf{0} \qquad \text{where} \qquad \tilde{g} \in \partial\left(\left\|\tilde{\beta}_\nu\right\|_1\right). \tag{B.3}$$

If $\hat{\beta}_\lambda = \tilde{\beta}_\nu$, then simply comparing (B.2) and (B.3) we have that $\lambda = 2\nu\|\tilde{\beta}_\nu\|_1$, and $\nu = (2\|\hat{\beta}_\lambda\|_1)^{-1}\lambda$.

Now for $\hat{\beta}_\lambda$ that satisfies (B.2), by plugging $\lambda = 2\nu\|\hat{\beta}_\lambda\|_1$, we have that $\hat{\beta}_\lambda$ satisfies (B.3), i.e., $\tilde{\beta}_\nu = \hat{\beta}_\lambda$ where $\lambda = 2\nu\|\hat{\beta}_\lambda\|_1$. Following the same argument, for $\tilde{\beta}_\nu$ that satisfies (B.3), we take $\nu = (2\|\tilde{\beta}_\nu\|_1)^{-1}\lambda$, and find that $\tilde{\beta}_\nu$ satisfies (B.2). This implies that $\hat{\beta}_\lambda = \tilde{\beta}_\nu$, where $\nu = (2\|\tilde{\beta}_\nu\|_1)^{-1}\lambda$. □

B.11 Fast rate in prediction error of the squared lasso

Recall the squared lasso estimate of β^*:

$$\check{\beta} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + 2\lambda\|\beta\|_1^2. \tag{B.4}$$

It is well known that the fast rate is built on the compatibility condition of the lasso problem. Let $\mathcal{S} = \text{supp}(\beta^*)$, i.e., the support of the true regression coefficient β^*, the compatibility condition of the squared lasso problem requires that for all $\mu \in \mathbb{R}^p$ such that $\|\mu_{\mathcal{S}^c}\|_1 - \sigma \le 3\|\mu_{\mathcal{S}}\|_1$,

$$\|\mu_{\mathcal{S}}\|_1 + \frac{1}{4}\sigma \le \sqrt{|\mathcal{S}|}\frac{\|\mathbf{X}\mu\|_2}{\sqrt{n}\phi_0}. \tag{B.5}$$

The following theorem establishes that the fast rate prediction error and an estimation error rate of $\check{\beta}$ in (B.4) can be attained with a value of λ that does not depend on any unknown parameters.

Theorem 47. *Suppose that each column $\mathbf{X}_j$ of the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has been scaled so that $\|\mathbf{X}_j\|_2^2 = n$ for all $j = 1, \ldots, p$, and $\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I_n\right)$. If compatibility condition* (B.5) *holds, then for any $L > 0$, the solution $\check{\beta}$ in* (B.4) *with tuning parameter*

$$\lambda = \left(\frac{2\log p + 2L}{n}\right)^{1/2} \tag{B.6}$$

attains the following estimation error rate and fast rate bound in prediction with probability greater than $1 - e^{-L}$*:*

$$\frac{1}{2n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 \leq \frac{64\max\left\{\|\beta^*\|_1, \sigma\right\}^2 |\mathcal{S}|\left(\log p + L\right)}{\phi_0^2 n};$$
$$\left\|\beta^* - \check{\beta}\right\|_1 \leq \frac{16\max\left\{\|\beta^*\|_1, \sigma\right\}|\mathcal{S}|}{\phi_0^2}\left(\frac{2\log p + 2L}{n}\right)^{1/2}.$$

Proof. First by the optimality of $\check{\beta}$, we have

$$\frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\check{\beta}\right\|_2^2 + 2\lambda\left\|\check{\beta}\right\|_1^2 \leq \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta^*\|_2^2 + 2\lambda\|\beta^*\|_1^2,$$

which implies that

$$\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 \leq \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 2\lambda\|\beta^*\|_1^2 - 2\lambda\left\|\check{\beta}\right\|_1^2. \tag{B.7}$$

The following proof is considered in two cases:

(1). When $\|\beta^*\|_1 \geq \sigma$: Note that $\|\cdot\|_1^2$ is convex and by chain rule, for any $g \in \partial(\|\beta^*\|_1)$,

$$\left\|\check{\beta}\right\|_1^2 - \|\beta^*\|_1^2 \geq 2\|\beta^*\|_1 g^T\left(\check{\beta} - \beta^*\right).$$

For $j \in \mathcal{S}$, we have that $g_j = \operatorname{sign}(\beta_j^*)$. For any $j \in \mathcal{S}^C$, we let

$$g_j = \operatorname{sign}\left(\check{\beta}_j - \beta_j^*\right) = \operatorname{sign}\left(\check{\beta}_j\right).$$

Then g is still a valid sub-differential of $\|\beta^*\|_1$. Moreover, conditional on the event

$$\mathcal{T} = \left\{\frac{1}{n}\|\mathbf{X}^T\varepsilon\|_\infty \leq \lambda\sigma\right\},$$

from (B.7) we have

$$\begin{aligned}
\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 &\leq \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 4\lambda\|\beta^*\|_1 g^T\left(\beta^* - \check{\beta}\right) \\
&= \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 4\lambda\|\beta^*\|_1 g_{\mathcal{S}}^T\left(\beta_{\mathcal{S}}^* - \check{\beta}_{\mathcal{S}}\right) + 4\lambda\|\beta^*\|_1 g_{\mathcal{S}^C}^T\left(\beta_{\mathcal{S}^C}^* - \check{\beta}_{\mathcal{S}^C}\right) \\
&= \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 4\lambda\|\beta^*\|_1 g_{\mathcal{S}}^T\left(\beta_{\mathcal{S}}^* - \check{\beta}_{\mathcal{S}}\right) - 4\lambda\|\beta^*\|_1\left\|\beta_{\mathcal{S}^C}^* - \check{\beta}_{\mathcal{S}^C}\right\|_1 \\
&\leq \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 4\lambda\|\beta^*\|_1\left\|\beta_{\mathcal{S}}^* - \check{\beta}_{\mathcal{S}}\right\|_1 - 4\lambda\|\beta^*\|_1\left\|\beta_{\mathcal{S}^C}^* - \check{\beta}_{\mathcal{S}^C}\right\|_1.
\end{aligned}$$

Since $\sigma \leq \|\beta^*\|_1$ and $\mathcal{T}$ holds, we have that $\left\|\mathbf{X}^T\varepsilon\right\|_\infty / n \leq \lambda\sigma \leq \lambda \|\beta^*\|_1$, and thus

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 &\leq 2\lambda\|\beta^*\|_1\left\|\beta^* - \check{\beta}\right\|_1 + 4\lambda\|\beta^*\|_1\left\|\beta^*_S - \check{\beta}_S\right\|_1 - 4\lambda\|\beta^*\|_1\left\|\beta^*_{S^C} - \check{\beta}_{S^C}\right\|_1 \\ &= 2\lambda\|\beta^*\|_1\left(3\left\|\beta^*_S - \check{\beta}_S\right\|_1 - \left\|\beta^*_{S^C} - \check{\beta}_{S^C}\right\|_1\right).\end{aligned}$$

This first implies that $3\left\|\beta^*_S - \check{\beta}_S\right\|_1 \geq \left\|\beta^*_{S^C} - \check{\beta}_{S^C}\right\|_1$, and that

$$\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\|\beta^*\|_1\left\|\beta^*_{S^C} - \check{\beta}_{S^C}\right\|_1 \leq 6\lambda\|\beta^*\|_1\left\|\beta^*_S - \check{\beta}_S\right\|_1.$$

Then by compatibility condition,

$$\begin{aligned}&\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\|\beta^*\|_1\left\|\beta^* - \check{\beta}\right\|_1 \\ =&\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\|\beta^*\|_1\left\|\beta^*_S - \check{\beta}_S\right\|_1 + 2\lambda\|\beta^*\|_1\left\|\beta^*_{S^C} - \check{\beta}_{S^C}\right\|_1 \\ \leq&8\lambda\|\beta^*\|_1\left\|\beta^*_S - \check{\beta}_S\right\|_1 \leq \frac{8\lambda\|\beta^*\|_1\sqrt{|S|}\left\|\mathbf{X}\beta^* - \mathbf{X}\check{\beta}\right\|_2}{\sqrt{n}\phi_0} \\ \leq&\frac{1}{2n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + \frac{32\|\beta^*\|_1^2\lambda^2|S|}{\phi_0^2}.\end{aligned} \tag{B.8}$$

(2). When $\|\beta^*\|_1 < \sigma$: We define $\gamma^* \in \mathbb{R}^p$ as

$$\gamma^*_j = \begin{cases}\beta^*_j + \frac{\sigma - \|\beta^*\|_1}{|S|} & \text{if } \beta^*_j > 0 \\ \beta^*_j - \frac{\sigma - \|\beta^*\|_1}{|S|} & \text{if } \beta^*_j < 0 \\ 0 & \text{if } \beta^*_j = 0.\end{cases}$$

It is easy to check that $\|\gamma^*\|_1 = \sigma$. Also (B.7) implies that

$$\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 \leq \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T\mathbf{X}^T\varepsilon + 2\lambda\left(\|\beta^*\|_1^2 - \|\gamma^*\|_1^2 + \|\gamma^*\|_1^2 - \left\|\check{\beta}\right\|_1^2\right).$$

Then we have that

$$\left\|\check{\beta}\right\|_1^2 - \|\gamma^*\|_1^2 \geq 2\|\gamma^*\|_1 g^T\left(\check{\beta} - \gamma^*\right)$$

holds for all $g \in \partial(\|\gamma^*\|_1)$, and it further implies that

$$\|\gamma^*\|_1^2 - \left\|\check{\beta}\right\|_1^2 \leq 2\|\gamma^*\|_1 g^T\left(\gamma^* - \check{\beta}\right) = 2\sigma g^T\left(\beta^* - \check{\beta}\right) + 2\sigma g^T\left(\gamma^* - \beta^*\right).$$

Note that any $g \in \partial(\|\gamma^*\|_1)$ is also a valid sub-differential of $\|\beta^*\|_1$, and

$$g^T(\gamma^* - \beta^*) = \sigma - \|\beta^*\|_1 .$$

Thus we have

$$\begin{aligned}
\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 &\le \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T \mathbf{X}^T\varepsilon + 2\lambda\left(\|\beta^*\|_1^2 - \sigma^2 + 2\sigma g^T\left(\beta^* - \check{\beta}\right) + 2\sigma^2 - 2\sigma\|\beta^*\|_1\right) \\
&= \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T \mathbf{X}^T\varepsilon + 4\lambda\sigma g^T\left(\beta^* - \check{\beta}\right) + 2\lambda\left(\sigma - \|\beta^*\|_1\right)^2 \\
&\le \frac{2}{n}\left(\check{\beta} - \beta^*\right)^T \mathbf{X}^T\varepsilon + 4\lambda\sigma g^T\left(\beta^* - \check{\beta}\right) + 2\lambda\sigma^2
\end{aligned}$$

Since γ^* and β^* have the same support, we can again choose $g_j = \operatorname{sign}(\check{\beta}_j)$ for $j \in S^c$. Conditional on the event $\mathcal{T}$, it follows that

$$\begin{aligned}
\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 &\le 2\lambda\sigma\left\|\check{\beta} - \beta^*\right\|_1 + 4\lambda\sigma\left\|\check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\right\|_1 - 4\lambda\sigma\left\|\check{\beta}_{\mathcal{S}^c} - \beta^*_{\mathcal{S}^c}\right\|_1 + 2\lambda\sigma^2 \\
&= 6\lambda\sigma\left\|\check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\right\|_1 - 2\lambda\sigma\left\|\check{\beta}_{\mathcal{S}^c} - \beta^*_{\mathcal{S}^c}\right\|_1 + 2\lambda\sigma^2.
\end{aligned}$$

This implies that $3\|\check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\| + \sigma \ge \|\check{\beta}_{\mathcal{S}^c} - \beta^*_{\mathcal{S}^c}\|$. And then by the compatibility condition (B.5),

$$\begin{aligned}
\frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\sigma\left\|\check{\beta} - \beta^*\right\|_1 &= \frac{1}{n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + 2\lambda\sigma\left\|\check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\right\|_1 + 2\lambda\sigma\left\|\check{\beta}_{\mathcal{S}^c} - \beta^*_{\mathcal{S}^c}\right\|_1 \\
&\le 8\lambda\sigma\left\|\check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\right\|_1 + 2\lambda\sigma^2 \\
&\le 8\lambda\sigma\sqrt{|\mathcal{S}|}\frac{\left\|\mathbf{X}\left(\check{\beta} - \beta^*\right)\right\|_2}{\sqrt{n\phi_0}} \\
&\le \frac{1}{2n}\left\|\mathbf{X}\check{\beta} - \mathbf{X}\beta^*\right\|_2^2 + \frac{32\lambda^2\sigma^2|\mathcal{S}|}{\phi_0^2}. \qquad \text{(B.9)}
\end{aligned}$$

By the proof of Corollary 4.3 in Giraud (2014), we have

$$\mathrm{P}\left\{\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty > \sigma\left(\frac{2\log p + 2L}{n}\right)^{1/2}\right\} \le e^{-L}.$$

Thus taking λ in (B.6), we have that

$$\mathrm{P}(\mathcal{T}^c) = \mathrm{P}\left(\frac{1}{n}\left\|\mathbf{X}^T\varepsilon\right\|_\infty > \lambda\sigma\right) \le e^{-L}.$$

And the results follow from (B.8) and (B.9). □

B.12 Additional results in numerical studies

We include in this section some additional results in the numerical studies in Section 3.5 and Section 3.6. In particular, Fig B.1 and Fig B.2 present the complementary results (in different simulation regimes) to Fig 3.1 and Fig 3.2 in the paper respectively, and Table B.1 shows the p-values of the paired t-tests and the Wilcoxon signed-rank tests of the difference of various methods outputs in Fig 3.1 in the paper. Finally, Table B.2 presents the mean and standard errors of $\mathrm{E}(\hat{\sigma}/\sigma)$ of various estimators in the real data example.

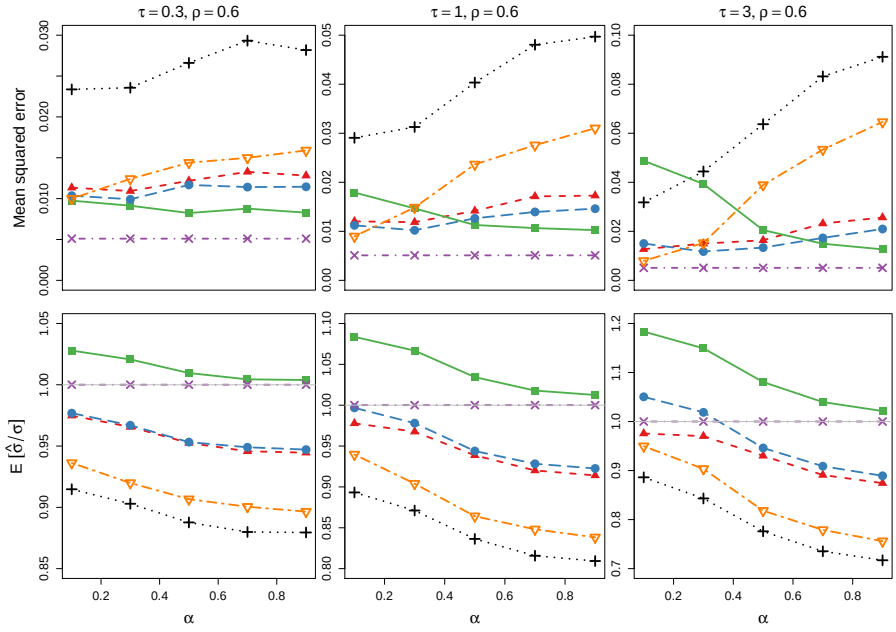

Figure B.1: Simulation results of various methods with regularization parameter selected using cross-validation. From left to right, column show the average (over 1000 repetitions) of the mean squared error (top panel) and $E(\hat{\sigma}/\sigma)$ (bottom panel) of various methods in three simulation settings. In each setting, we fix model sparsity (α) and correlations among features (ρ), and let signal-to-noise ratio(as expressed in τ) change. Line styles and their corresponding methods: + for naive, ▲ for $\hat{\sigma}_R^2$, ▽ for the square-root/scaled lasso, ▬ for the natural lasso, ● for the organic lasso, × for the oracle.

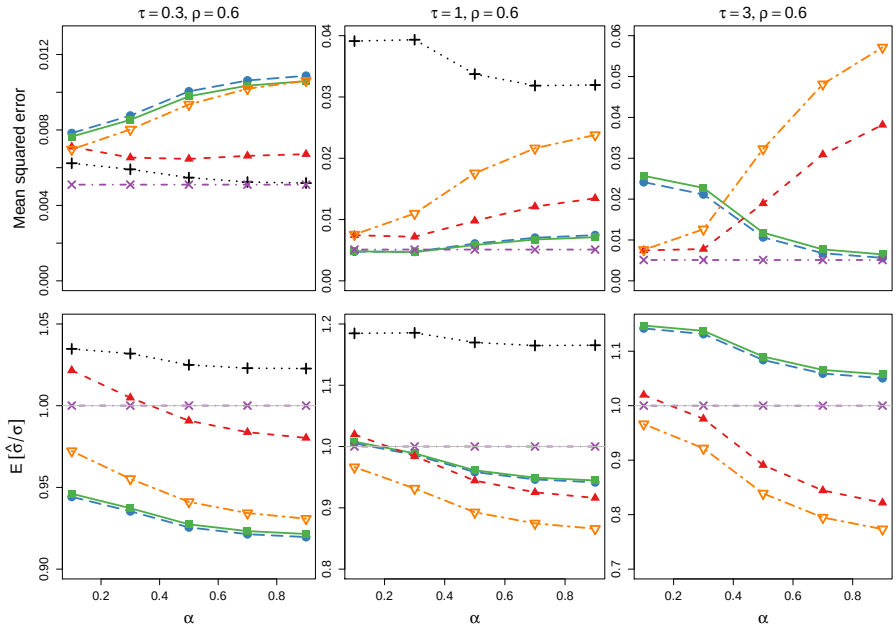

Figure B.2: Simulation results of various methods with pre-specified regularization parameter values. From left to right, column show the average (over 1000 repetitions) of the mean squared error (top panel) and $E(\hat{\sigma}/\sigma)$ (bottom panel) of various methods in three simulation settings. In each setting, we fix model sparsity (α) and correlations among features (ρ), and let signal-to-noise ratio(as expressed in τ) change. Line styles and their corresponding methods: + for organic (λ_0), for organic (λ_2), for organic (λ_3), for scaled(1), for scaled (2), for the oracle.

Table B.1: p-values for testing the difference of various methods outputs

	natural vs. organic	$\hat{\sigma}_R^2$ vs. organic	$\hat{\sigma}_R^2$ vs. natural
$\alpha = 0.1, \rho = 0.3, \tau = 1$	0.00 (0.00)	0.07 (0.00)	0.00 (0.00)
$\alpha = 0.3, \rho = 0.3, \tau = 1$	0.00 (0.00)	0.19 (0.25)	0.00 (0.00)
$\alpha = 0.5, \rho = 0.3, \tau = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.7, \rho = 0.3, \tau = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.9, \rho = 0.3, \tau = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.1, \rho = 0.6, \tau = 1$	0.00 (0.00)	0.08 (0.01)	0.00 (0.00)
$\alpha = 0.3, \rho = 0.6, \tau = 1$	0.00 (0.00)	0.00 (0.14)	0.00 (0.00)
$\alpha = 0.5, \rho = 0.6, \tau = 1$	0.05 (0.10)	0.01 (0.00)	0.00 (0.00)
$\alpha = 0.7, \rho = 0.6, \tau = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.9, \rho = 0.6, \tau = 1$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.1, \rho = 0.9, \tau = 1$	0.06 (0.32)	0.00 (0.03)	0.00 (0.12)
$\alpha = 0.3, \rho = 0.9, \tau = 1$	0.96 (0.02)	0.00 (0.07)	0.00 (0.00)
$\alpha = 0.5, \rho = 0.9, \tau = 1$	0.03 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.7, \rho = 0.9, \tau = 1$	0.44 (0.00)	0.00 (0.00)	0.00 (0.00)
$\alpha = 0.9, \rho = 0.9, \tau = 1$	0.20 (0.00)	0.00 (0.01)	0.00 (0.00)

In each simulation setting, as characterized by a (α, ρ, τ) triplet, we report p-values of the (two-sided) paired t-tests and the Wilcoxon signed-rank tests (shown in parentheses) for testing the null hypothesis that the output of each pair of methods are the same.

Table B.2: $E(\hat{\sigma}/\sigma)$ in MSD dataset

n	20	40	60	80	100	120
naive	80.1 (1.1)	94.2 (0.9)	95.8 (0.7)	96.4 (0.6)	97.9 (0.5)	96.7 (0.5)
$\hat{\sigma}_R^2$	90.0 (1.0)	100.4 (0.8)	101.7 (0.6)	102.3 (0.5)	103.3 (0.5)	102.4 (0.4)
natural	94.0 (0.9)	103.3 (0.7)	105.5 (0.6)	106.0 (0.5)	107.0 (0.4)	106.6 (0.4)
organic	86.8 (0.8)	97.6 (0.6)	99.9 (0.5)	100.9 (0.4)	101.7 (0.4)	101.8 (0.4)
scaled(1)	106.1 (0.8)	109.3 (0.6)	111.2 (0.5)	111.2 (0.4)	111.7 (0.4)	111.8 (0.4)
scaled(2)	88.5 (0.8)	99.0 (0.6)	102.9 (0.5)	104.4 (0.5)	105.1 (0.4)	105.5 (0.3)
organic(λ_2)	89.7 (0.7)	94.7 (0.5)	97.6 (0.4)	98.3 (0.4)	99.2 (0.3)	99.7 (0.3)
organic(λ_3)	92.0 (0.7)	97.3 (0.6)	100.1 (0.4)	100.7 (0.4)	101.6 (0.4)	102.0 (0.3)

Mean and standard errors (over 1000 replications) of $E(\hat{\sigma}/\sigma)$ of various methods we considered in Section 3.5. Each entry of the method output is multiplied by 100 to convey information more compactly.

APPENDIX C

APPENDIX OF CHAPTER 4

Lemma 48. *Under Assumption **A1**, the following inequalities hold for some constants* $C_1, C_2, C_3, C_4 > 0$*:*

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{ij}\mathbf{X}_{ik} - \mathrm{E}(Z_jX_k)\right| > \varepsilon\right) \le C_1\exp(-C_2n^{\frac{2}{3}}\varepsilon)$$
$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{ij}\mathbf{Z}_{ik} - \mathrm{E}(Z_jZ_k)\right| > \varepsilon\right) \le C_3\exp(-C_4n^{\frac{1}{2}}\varepsilon).$$

Furthermore, the results above hold when $\mathbf{Z}$ *and* Z *are replaced with* $\mathbf{W}$ *and* W *respectively.*

Proof. For each i, let $A_i = \mathbf{Z}_{ij}\mathbf{X}_{ik} = \mathbf{X}_{il}\mathbf{X}_{ih}\mathbf{X}_{ik}$ for some $1 \le l \le h \le p$. Under the sub-Gaussian assumption, there exists some constant $C > 0$ such that $\mathrm{E}\exp(C\mathbf{X}_{ij}^2) \le 2$, $\mathrm{E}\exp(C\mathbf{X}_{il}^2) \le 2$, and $\mathrm{E}\exp(C\mathbf{X}_{ih}^2) \le 2$. Then

$$\begin{aligned}\mathrm{E}\exp\left(C|\mathbf{Z}_{ij}\mathbf{X}_{ik}|^{\frac{2}{3}}\right) &\le \mathrm{E}\exp\left(\frac{C\mathbf{X}_{il}^2 + C\mathbf{X}_{ih}^2 + C\mathbf{X}_{ik}^2}{3}\right)\\ &\le \frac{1}{3}\mathrm{E}\exp\left(C\mathbf{X}_{il}^2\right) + \frac{1}{3}\mathrm{E}\exp\left(C\mathbf{X}_{ih}^2\right) + \frac{1}{3}\mathrm{E}\exp\left(C\mathbf{X}_{ik}^2\right) \le 2.\end{aligned}$$

We've shown that the independent random variables A_i has mean zero and $\mathrm{E}(e^{C|A_i|^{2/3}}) \le 2$, and the first inequality follows from Lemma B.4 in Hao & Zhang (2014).

Next, suppose $Z_j = X_aX_b$ and $Z_k = X_lX_h$ for some $1 \le a \le b \le p$, and $1 \le l \le h \le p$. Consider $B_i = \mathbf{X}_{ia}\mathbf{X}_{ib}\mathbf{X}_{il}\mathbf{X}_{ih} - \mathrm{E}(Z_jZ_k)$. Then $\mathrm{E}(B_i) = 0$, and B_i are independent.

It is left to show that there exist some constants $T_0, A_0 > 0$ such that $\mathrm{E}(e^{T_0|B_i|^{1/2}}) \le A_0$. It can be check that Z_jZ_k is a sub-exponential random variable.

Thus there exists some constant $K > 0$ such that $\mathrm{E}|Z_j Z_k| \leq K^2$. And we have

$$\begin{aligned}
\mathrm{E}\left\{\exp\left(C|B_i|^{\frac{1}{2}}\right)\right\} &= \mathrm{E}\left\{\exp\left(C|\mathbf{X}_{ia}\mathbf{X}_{ib}\mathbf{X}_{ih}\mathbf{X}_{il} - \mathrm{E}(Z_j Z_k)|^{\frac{1}{2}}\right)\right\} \\
&\leq \mathrm{E}\left\{\exp\left(C|\mathbf{X}_{ia}\mathbf{X}_{ib}\mathbf{X}_{ih}\mathbf{X}_{il}|^{\frac{1}{2}} + C|\mathrm{E}(Z_j Z_k)|^{\frac{1}{2}}\right)\right\} \\
&\leq \exp(CK)\mathrm{E}\left\{\exp\left(C\frac{\mathbf{X}_{ia}^2 + \mathbf{X}_{ib}^2 + \mathbf{X}_{il}^2 + \mathbf{X}_{ih}^2}{4}\right)\right\} \\
&\leq \frac{\exp(CK)}{4}\mathrm{E}\left\{\exp\left(C\mathbf{X}_{ia}^2\right) + \exp\left(\mathbf{X}_{ib}^2\right) + \exp\left(\mathbf{X}_{il}^2\right) + \exp\left(\mathbf{X}_{ih}^2\right)\right\} \\
&\leq 2\exp(CK).
\end{aligned}$$

□

C.1 Proof of Theorem 20

We follow the analysis in Barut et al. (2016) and Fan et al. (2016). First we let the vector $\mathbf{1}_n$ stands for a vector of n ones, and $\mathbf{C}_n = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n$ is the centering matrix. We consider

$$\omega_j = \frac{\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{r}}{\sqrt{\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{Z}_j}} = \widehat{\mathrm{sd}}(\mathbf{r})\widehat{\mathrm{corr}}\left(\mathbf{Z}_j, \mathbf{r}\right), \tag{C.1}$$

and the corresponding population quantity

$$\omega_j^* = \frac{\mathrm{Cov}\left(Z_j, W\right)\gamma^*}{\sqrt{\Psi_{jj}}} = \frac{\Omega_j^T\gamma^*}{\sqrt{\Psi_{jj}}}. \tag{C.2}$$

We first show that ω_j^* is useful in representing interaction variables $j \in \mathcal{I}$, and furthermore that ω_j converges to ω_j^*. As a result, we can use ω_j, which is computable, as a noisy proxy for ω_j^* to determine whether j is in $\mathcal{I}$. We formally present it as the following lemma

Lemma 49. *Under Assumptions **A1**, **A2** and **A3**, if* $\sqrt{2\max_j \Psi_{jj}}n^{-\tau} \leq 6^{-1}C_\kappa n^{-\kappa}$ *and* $\xi + 2\xi_1 + 2\kappa < \frac{1}{2}$*, then*

$$\mathrm{P}\left(\max_{1\leq j\leq q}|\omega_j - \omega_j^*| \leq \frac{1}{2}C_\kappa n^{-\kappa}\right) \geq 1 - c_1\exp(-c_2 n^\xi) \tag{C.3}$$

holds for some constants $c_1, c_2 > 0$*;*

Proof. We start by rewriting (C.1) and (C.2) as

$$\omega_j = \frac{A_j}{\sqrt{B_j}} \qquad \omega_j^* = \frac{\Omega_j^T\gamma^*}{\sqrt{\Psi_{jj}}},$$

where $A_j = \frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{r}$ and $B_j = \frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{Z}_j$. First note that

$$\begin{aligned}
&\max_{1\leq j\leq q}|A_j - \Omega_j^T\gamma^*| = \max_{1\leq j\leq q}\left|\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\left(\mathbf{W}\gamma^* + \mathbf{X}\theta^* - \mathbf{X}\hat{\theta} + \boldsymbol{\varepsilon}\right) + \Omega_j^T\gamma^*\right| \\
&= \max_{1\leq j\leq q}\left|\left(\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{W} - \Omega_j^T\right)\gamma^* + \frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{X}(\theta^* - \hat{\theta}) + \frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\boldsymbol{\varepsilon}\right| \\
&\leq \max_{1\leq j\leq q}\max_{1\leq k\leq q}\left|\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{W}_k - \Omega_{jk}\right|\|\gamma^*\|_1 + \max_{1\leq j\leq q}\frac{1}{n}\left\|\mathbf{C}_n\mathbf{Z}_j\right\|_2\left\|\mathbf{X}\theta^* - \mathbf{X}\hat{\theta}\right\|_2 + \max_{1\leq j\leq q}\left|\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\boldsymbol{\varepsilon}\right|.
\end{aligned}$$

First note that $\|\gamma^*\|_1 \leq A_1 n^{\xi_1}$. By Lemma 48, for any $x > 0$, we have

$$\begin{aligned}
\mathrm{P}\left\{\max_{1\leq j\leq q}\max_{1\leq k\leq q}\left|\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{W}_k - \Omega_{jk}\right|\|\gamma^*\|_1 > xA_1n^{\xi_1}\right\} &\leq C_3q^2\exp\left(-2C_4n^{\frac{1}{2}}x^2\right) \\
&= C_3\exp\left\{2n^\xi\left(1 - C_4n^{\frac{1}{2}-\xi}x^2\right)\right\}.
\end{aligned}$$

Take $x = C_\kappa(6A_1)^{-1}n^{-\xi_1-\kappa}$, then

$$\begin{aligned}
\mathrm{P}\left\{\max_{1\leq j\leq q}\max_{1\leq k\leq q}\left|\frac{1}{n}\mathbf{Z}_j^T\mathbf{C}_n\mathbf{W}_k - \Omega_{jk}\right|\|\gamma^*\|_1 > \frac{C_\kappa n^{-\kappa}}{6}\right\} &\leq C_3\exp\left\{2n^\xi\left(1 - \frac{C_4C_\kappa}{6A_1}n^{\frac{1}{2}-\xi-2\xi_1-2\kappa}\right)\right\} \\
&\leq C_3\exp\left(-\tilde{C}_1n^\xi\right).
\end{aligned} \tag{C.4}$$

Consider the event $\mathcal{E}_j = \{\|\mathbf{C}_n\mathbf{Z}_j\|_2^2 \leq 2\Psi_{jj}n\}$. Lemma 48 implies that

$$\mathrm{P}\left(\mathcal{E}_j^C\right) = \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n(\mathbf{C}_n\mathbf{Z})_{ij}^2 > 2\Psi_{jj}\right) \leq \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^n(\mathbf{C}_n\mathbf{Z}_{ij})^2 - \Psi_{jj}\right| > 2\Psi_{jj} - \Psi_{jj}\right) \leq C_3\exp\left(-C_4\Psi_{jj}n^{\frac{1}{2}}\right).$$

Then conditional on event (4.6) and by the condition that $\sqrt{2\max_j \Psi_{jj}}n^{-\tau} \leq 6^{-1}C_\kappa n^{-\kappa}$, we have

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leq j\leq q}\frac{1}{n}\left\|\mathbf{C}_n\mathbf{Z}_j\right\|_2\left\|\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*\right\|_2 > \frac{C_\kappa}{6}n^{-\kappa}\right)\\
\leq&\mathrm{P}\left(\max_{1\leq j\leq q}\frac{1}{n}\left\|\mathbf{C}_n\mathbf{Z}_j\right\|_2\left\|\mathbf{X}\hat{\theta}-\mathbf{X}\theta^*\right\|_2 > \sqrt{2\max_j \Psi_{jj}}n^{-\tau}\right)\\
\leq&C_3\exp\left\{n^\xi\left(1-C_4\max_j\Psi_{jj}n^{\frac{1}{2}-\xi}\right)\right\} \leq C_3\exp(-\tilde{C}_2 n^\xi). \qquad \text{(C.5)}
\end{aligned}$$

Recall that ε_i are independent sub-Gaussian random variables, i.e., $\mathrm{E}\exp(\varepsilon_i^2/K^2) \leq 2$ for some $K^2 > 0$. Conditioning on that $\mathbf{Z}_j = \tilde{\mathbf{Z}}_j$, we have $n^{-1}\tilde{\mathbf{Z}}_j^T\boldsymbol{\varepsilon}$ is a sub-Gaussian random variable. Then on $\mathcal{E}_j$, by the following Hoeffding-type inequality (Vershynin 2010) we have

$$\mathrm{P}\left(\frac{1}{n}\left|\mathbf{Z}_j^T\mathbf{C}_n\boldsymbol{\varepsilon}\right| > x\middle|\mathbf{Z}_j=\tilde{\mathbf{Z}}_j\right) \leq \exp\left(1-\frac{C_7x^2n^2}{K^2\|\mathbf{C}_n\tilde{\mathbf{Z}}_j\|_2^2}\right) \leq \exp\left(1-\frac{C_7nx^2}{2\Psi_{jj}K^2}\right),$$

for some constant $C_7 > 0$. So take $x = C_\kappa 6^{-1}n^{-\kappa}$ and use a union bound, we have

$$\begin{aligned}
\mathrm{P}\left(\max_{1\leq j\leq q}\frac{1}{n}\left|\mathbf{Z}_j^T\mathbf{C}_n\boldsymbol{\varepsilon}\right| > \frac{C_\kappa n^{-\kappa}}{6}\right) &\leq C_3\exp\left\{n^\xi\left(1-C_4\Psi_{jj}n^{\frac{1}{2}-\xi}\right)\right\} + \exp\left\{n^\xi\left(1+n^{-\xi}-\frac{C_7C_\kappa^2n^{1-2\kappa-\xi}}{72\Psi_{jj}K^2}\right)\right\}\\
&\leq C_8\exp\left(-C_9n^\xi\right). \qquad \text{(C.6)}
\end{aligned}$$

Combining (C.4), (C.5), and (C.6), a union bound implies that

$$\mathrm{P}\left(\max_{1\leq j\leq q}|A_j-\Omega_j^T\gamma^*| > \frac{1}{2}C_\kappa n^{-\kappa}\right) \leq c_1\exp\left(-c_2n^\xi\right) \qquad \text{(C.7)}$$

for some constants $c_1, c_2 > 0$.

By Lemma 48,

$$\mathrm{P}\left(\max_{1\leq j\leq p}\left|B_j-\Psi_{jj}\right| > \frac{1}{2}C_\kappa n^{-\kappa}\right) \leq C_3\exp\left(n^\xi-\frac{C_4C_\kappa}{2}n^{\frac{1}{2}-\kappa}\right) \leq c_3\exp\left(-c_4n^\xi\right) \qquad \text{(C.8)}$$

for some constants $c_3, c_4 > 0$. So combining (C.7) and (C.8), Lemma 10 and Lemma 12 in Fan et al. (2016) imply the first part of the theorem.

Now consider the event

$$\mathcal{E} = \left\{ \max_{j \in \mathcal{I}} |\omega_j - \omega_j^*| \leq \frac{1}{2} C_\kappa n^{-\kappa} \right\}.$$

Suppose $\mathcal{E}$ holds, for any $j \in \mathcal{I}$, by Assumption **A4**,

$$|\omega_j| \geq |\omega_j^*| - |\omega_j^* - \omega_j| > C_\kappa n^{-\kappa},$$

which implies that $j \in \hat{\mathcal{I}}_\eta$ with $\eta = C_\kappa n^{-\kappa}$. Thus

$$\mathrm{P}\left(\mathcal{I} \subseteq \hat{\mathcal{I}}_\eta\right) \geq \mathrm{P}\left(\mathcal{E}\right) = 1 - \mathrm{P}\left(\max_{j \in \mathcal{I}} |\omega_j - \omega_j^*| > \frac{1}{2} C_\kappa n^{-\kappa} \right).$$

□

To show Theorem 20, we first give an upper bound on $\sum_{j=1}^q \omega_j^{*2}$. First note that

$$\sum_{j=1}^q \omega_j^{*2} = \sum_{j=1}^p \Psi_{jj}^{-1} (\Omega_j^T \gamma^*)^2 = \left\| \mathrm{diag}(\Psi)^{-1/2} \Omega \gamma^* \right\|_2^2 \leq \lambda_{\max} \left(\mathrm{diag}(\Psi)^{-1/2} \Omega \right) \gamma^{*T} \Omega \gamma^*,$$

and that

$$\mathrm{Var}(Y) = \beta^{*T} \Sigma \beta^* + \gamma^{*T} \Omega \gamma^* + \sigma^2 \geq \gamma^{*T} \Omega \gamma^*,$$

which together imply that $\sum_{j=1}^q \omega_j^{*2} \leq \lambda_{\max} \left(\mathrm{diag}(\Psi)^{-1/2} \Omega \right) \mathrm{Var}(Y)$. Consider the set $\tilde{\mathcal{I}} = \{ j : |\omega_j^*| > 2^{-1} C_\kappa n^{-\kappa} \}$. Conditional on $\mathcal{E}$, for any $j \in \hat{\mathcal{I}}_\eta$ with $\eta = C_\kappa n^{-\kappa}$, we have that $|\omega_j^*| \geq |\omega_j| - |\omega_j - \omega_j^*| > 2^{-1} C_\kappa n^{-\kappa}$, which implies that $j \in \tilde{\mathcal{I}}$. Thus

$$|\hat{\mathcal{I}}_\eta| \leq |\tilde{\mathcal{I}}| \leq \frac{4 \lambda_{\max} \left(\mathrm{diag}(\Psi)^{-1/2} \Omega \right) \mathrm{Var}(Y) n^{2\kappa}}{C_\kappa^2}.$$

C.2 Proof of Theorem 21

The basic inequality implies that

$$\frac{1}{2n} \left\| \mathbf{y} - \mathbf{X}\hat{\theta} - \hat{\mathbf{W}}_{\hat{\mathcal{I}}} \hat{\eta} \right\|_2^2 + \lambda \left(\|\hat{\theta}\|_1 + \|\hat{\eta}\|_1 \right) \leq \frac{1}{2n} \left\| \mathbf{y} - \mathbf{X}\theta^* - \hat{\mathbf{W}}_{\hat{\mathcal{I}}} \eta^* \right\|_2^2 + \lambda \left(\|\theta^*\|_1 + \|\eta^*\|_1 \right)$$

for any pair of θ^* and η^*. For $\theta^* = \beta^* + \vartheta^*$ in (4.3), we have

$$\begin{aligned}
&\frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\hat{\theta} - \hat{\mathbf{W}}_{\hat{I}}\hat{\eta}\right\|_2^2 + \lambda\left(\|\hat{\theta}\|_1 + \|\hat{\eta}\|_1\right) = \frac{1}{2n}\left\|\mathbf{X}\left(\theta^* - \hat{\theta}\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\hat{\eta} + \boldsymbol{\varepsilon}\right\|_2^2 + \lambda\left(\|\hat{\theta}\|_1 + \|\hat{\eta}\|_1\right) \\
&\leq \frac{1}{2n}\left\|\boldsymbol{\varepsilon} + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
&= \frac{1}{2n}\|\boldsymbol{\varepsilon}\|_2^2 + \frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \frac{1}{n}\boldsymbol{\varepsilon}^T\left(\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right) + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\frac{1}{2n}\left\|\mathbf{X}\left(\theta^* - \hat{\theta}\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\hat{\eta}\right\|_2^2 + \lambda\left(\|\hat{\theta}\|_1 + \|\hat{\eta}\|_1\right) \\
\leq &\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}\left(\hat{\theta} - \theta^*\right) + \frac{1}{n}\boldsymbol{\varepsilon}^T\left(\hat{\mathbf{W}}_{\hat{I}}\hat{\eta} - \mathbf{W}\gamma^*\right) + \frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \frac{1}{n}\boldsymbol{\varepsilon}^T\left(\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right) + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
= &\frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{X}\left(\hat{\theta} - \theta^*\right) + \frac{1}{n}\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_{\hat{I}}\left(\hat{\eta} - \eta^*\right) + \frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
\leq &\frac{1}{n}\max_j|\boldsymbol{\varepsilon}^T\mathbf{X}_j|\left(\|\hat{\theta}\|_1 + \|\theta^*\|_1\right) + \frac{1}{n}\max_{j\in\hat{I}}|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j|\|\hat{\eta} - \eta^*\|_1 + \frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
\leq &\frac{1}{n}\max_j|\boldsymbol{\varepsilon}^T\mathbf{X}_j|\left(\|\hat{\theta}\|_1 + \|\theta^*\|_1\right) + \frac{1}{n}\max_{j\in\hat{I}}|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j|\left(\|\hat{\eta}\|_1 + \|\eta^*\|_1\right) + \frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + \lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right).
\end{aligned}$$

On the event

$$\mathcal{T}_1 = \left\{\frac{1}{n}\max_{1\leq j\leq p}|\boldsymbol{\varepsilon}^T\mathbf{X}_j| \leq \lambda\right\} \quad \cap \quad \mathcal{T}_2 = \left\{\frac{1}{n}\max_{j\in\hat{I}}|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j| \leq \lambda\right\}, \tag{C.9}$$

we have

$$\begin{aligned}
&\frac{1}{2n}\left\|\mathbf{X}\left(\hat{\theta} - \theta^*\right) + \mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\hat{\eta}\right\|_2^2 \\
\leq &\frac{1}{2n}\left\|\mathbf{W}\gamma^* - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + 2\lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
= &\frac{1}{2n}\left\|\mathbf{W}_{\hat{I}^C}\gamma^*_{\hat{I}^C} + \mathbf{W}_{\hat{I}}\gamma^*_{\hat{I}} - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + 2\lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
\leq &\frac{1}{n}\left\|\mathbf{W}_{\hat{I}^C}\gamma^*_{\hat{I}^C}\right\|_2^2 + \frac{1}{n}\left\|\mathbf{W}_{\hat{I}}\gamma^*_{\hat{I}} - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + 2\lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right) \\
\leq &\frac{|\hat{I}^C \cap \mathrm{supp}(\gamma^*)|}{n}\sum_{j\in\hat{I}^C\cap\mathrm{supp}(\gamma^*)}\left\|\mathbf{W}_j\gamma^*_j\right\|_2^2 + \frac{1}{n}\left\|\mathbf{W}_{\hat{I}}\gamma^*_{\hat{I}} - \hat{\mathbf{W}}_{\hat{I}}\eta^*\right\|_2^2 + 2\lambda\left(\|\theta^*\|_1 + \|\eta^*\|_1\right).
\end{aligned}$$

where the two last inequalities hold because $\|\sum_{k\in\mathcal{A}}a_k\|_2^2 = \sum_{k\in\mathcal{A}}\|a_k\|_2^2 + 2\sum_{j<\ell}a_j^Ta_\ell \leq \sum_{k\in\mathcal{A}}\|a_k\|_2^2 + 2\sum_{j<\ell}\|a_j\|_2\|a_\ell\|_2 \leq |\mathcal{A}|\sum_{k\in\mathcal{A}}\|a_k\|_2^2$ for any set of vectors $\{a_k\}_{k\in\mathcal{A}}$. Let $\eta^* = \gamma^*_{\hat{I}}$, then Theorem 21 follows.

C.3 Proof of Corollary 22

We characterize the value of λ such that (4.9) holds. First using the proof of Lemma 50, we have that for any $\lambda > 0$, the following bound holds for some constants $c_1, c_2, c_3 > 0$:

$$\mathrm{P}\left(\max_{1\le j\le p}\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{X}_j\right| > \lambda\right) \le p\exp\left(1 - \frac{c_1 n\lambda^2}{2\sigma^2\max_j \Sigma_{jj}}\right) + c_3 p\exp(-c_2 n).$$

Take λ as in (4.11) where $c = 2c_1^{-1/2}$, we have

$$\begin{aligned}\mathrm{P}(\mathcal{T}_1^C) &\le \exp(1-\log p) + c_3\exp(\log p - c_2 n)\\ &\le \exp(1-\log p) + c_3\exp\left\{-n\left(c_2 - n^{-1}\log p\right)\right\} \le \exp(1-\log p) + C_1\exp(-C_2 n)\end{aligned} \tag{C.10}$$

for some $C_1, C_2 > 0$.

For $\hat{\mathbf{W}} = \mathbf{W}$ and each $j \in \hat{\mathcal{I}}$, $\boldsymbol{\varepsilon}$ and $\mathbf{W}_j$ are independent, and $\mathrm{E}(\boldsymbol{\varepsilon}^T\mathbf{W}_j) = \mathrm{E}(\boldsymbol{\varepsilon})^T\mathrm{E}(\mathbf{W}_j) = 0$. So conditional on $\mathbf{W}_j$, $\boldsymbol{\varepsilon}^T\mathbf{W}_j$ follows a sub-Gaussian distribution with mean zero and variance $\sigma^2\|\mathbf{W}_j\|_2^2$. From a Hoeffding-type inequality we have

$$\mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\middle|\mathbf{W}_j\right) \le \exp\left(1 - \frac{c_4\lambda^2 n^2}{\sigma^2\|\mathbf{W}_j\|_2^2}\right).$$

Consider the event $\mathcal{E}_j = \left[\|\mathbf{W}_j\|_2^2 \le n\{x + \mathrm{E}(W_j^2)\}\right]$ for any $x > 0$, we have

$$\begin{aligned}\mathrm{P}(\mathcal{E}_j^C) &= \mathrm{P}\left(\frac{1}{n}\|\mathbf{W}_j\|_2^2 > x + \mathrm{E}(W_j^2)\right)\\ &\le \mathrm{P}\left(\left|\frac{1}{n}\|\mathbf{W}_j\|_2^2 - \mathrm{E}(W_j^2)\right| > x\right)\\ &\le c_5\exp(-c_6 n^{1/2}x),\end{aligned}$$

where the last inequality follows from Lemma 48, and that

$$\mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\middle|\mathcal{E}_j\right) \le \exp\left(1 - \frac{c_4\lambda^2 n}{\sigma^2\{x + \mathrm{E}(W_j^2)\}}\right).$$

Therefore,

$$\begin{aligned}\mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\right) &\le \mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\middle|\mathcal{E}_j\right)\mathrm{P}(\mathcal{E}_j) + \mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\middle|\mathcal{E}_j^C\right)\mathrm{P}(\mathcal{E}_j^C) \\ &\le \exp\left[1 - \frac{c_4\lambda^2 n}{\sigma^2\{x + \mathrm{E}(W_j^2)\}}\right] + c_5\exp(-c_6 n^{1/2}x).\end{aligned}$$

By a union bound and taking $x = \mathrm{E}(W_j^2)$

$$\mathrm{P}\left(\frac{1}{n}\max_{j\in\hat{\mathcal{I}}}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\right) \le |\hat{\mathcal{I}}|\exp\left\{1 - \frac{c_4\lambda^2 n}{2\sigma^2\mathrm{E}(W_j^2)}\right\} + c_5|\hat{\mathcal{I}}|\exp\left\{-c_6 n^{1/2}\max_j \mathrm{E}(W_j^2)\right\}.$$

Conditional on Theorem 20 holds, we have $|\hat{\mathcal{I}}| \le 4C_\kappa^{-2}\lambda_{\max}\left(\mathrm{diag}(\Omega)^{-1/2}\Omega\right)\mathrm{Var}(Y)n^{2\kappa} := Dn^{2\kappa}$. Then take λ as in (4.11), we have

$$\begin{aligned}&\mathrm{P}\left(\frac{1}{n}\max_{j\in\hat{\mathcal{I}}}\left|\boldsymbol{\varepsilon}^T\mathbf{W}_j\right| > \lambda\right) \\ &\le \exp\left(1 + 2\kappa\log n + \log D - \frac{4c_1^{-1}c_4\max_j\Sigma_{jj}\log p}{2\mathrm{E}(W_j^2)}\right) + c_5\exp\left\{2\kappa\log n + \log D - c_6 n^{1/2}\max_j\mathrm{E}(W_j^2)\right\} \\ &\le \exp\left\{-\log p\left(\frac{4c_1^{-1}c_4\max_j\Sigma_{jj}}{2\mathrm{E}(W_j^2)} - \frac{1 + 2\kappa\log n + \log D}{\log p}\right)\right\} \\ &\quad + c_5\exp\left\{-n^{1/2}\left(c_6\max_j\mathrm{E}(W_j^2) - \frac{2\kappa\log n + \log D}{n^{1/2}}\right)\right\} \\ &\le \exp(-C_3\log p) + C_4\exp(-C_5 n^{1/2}) \qquad \text{(C.11)}\end{aligned}$$

for some $C_3, C_4, C_5 > 0$ where the last inequality holds if $n^{2\kappa} = o(p)$. The probability bound in Corollary 22 then follows from (C.10) and (C.11) by taking $K_1 = e + 1$, $K_2 = \min(C_3, 1)$, $K_3 = C_1 + C_4$, and $K_4 = \min(C_2, C_5)$.

Finally, on the event $\mathcal{E}_j$ with $x = \mathrm{E}(W_j^2)$, with the same probability we have $\frac{1}{n}\|\mathbf{W}_j\gamma^*\|_2^2 \le 2\mathrm{E}(W_j^2)\gamma_j^{*2}$. Given that $\mathcal{I} \subseteq \hat{\mathcal{I}}$, where $\mathcal{I}$ is defined as in (4.7), we have $\hat{\mathcal{I}}^C \subseteq \mathcal{I}^C$, and thus $\frac{1}{n}\|\mathbf{W}_j\gamma^*\|_2^2 \le 2\mathrm{E}(W_j^2)\gamma_j^{*2} \le 2\alpha$. And the result follows from (4.10).

C.4 Proof of Corollary 23

For each $j \in \hat{\mathcal{I}}$, $\boldsymbol{\varepsilon}$ and $\hat{\mathbf{W}}_j$ are independent, and $\mathrm{E}(\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j) = \mathrm{E}(\boldsymbol{\varepsilon})^T\mathrm{E}(\hat{\mathbf{W}}_j) = 0$. So conditional on $\hat{\mathbf{W}}_j$, $\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j$ follows a sub-Gaussian distribution with mean zero and variance $\sigma^2\|\hat{\mathbf{W}}_j\|_2^2$. From a Hoeffding-type inequality we have

$$\mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j\right| > \lambda\middle|\hat{\mathbf{W}}_j\right) \le \exp\left(1 - \frac{c_4\lambda^2 n^2}{\sigma^2\|\hat{\mathbf{W}}_j\|_2^2}\right).$$

Consider the event $\hat{\mathcal{E}}_j = \left[\|\hat{\mathbf{W}}_j\|_2^2 \le C\|\mathbf{W}_j\|_2^2 \le Cn\{x + \mathrm{E}(W_j^2)\}\right]$ for any $x > 0$, then the rest of the proof is the same as in Section C.3.

C.5 Proof of Corollary 24

Recall that $\mathbf{W}_j = \mathbf{Z}_j - \mathbf{X}_j\phi^{(j)}$, where $\phi^{(j)} = \Sigma^{-1}\Phi_j$. From the basic inequality of (4.5),

$$\frac{1}{2n}\left\|\mathbf{Z}_j - \mathbf{X}\hat{\phi}^{(j)}\right\|_2^2 + \nu\|\hat{\phi}^{(j)}\|_1 \le \frac{1}{2n}\left\|\mathbf{Z}_j - \mathbf{X}\phi^{(j)}\right\|_2^2 + \nu\|\phi^{(j)}\|_1,$$

which is equivalent to

$$\frac{1}{2n}\left\|\mathbf{W}_j + \mathbf{X}\left(\phi^{(j)} - \hat{\phi}^{(j)}\right)\right\|_2^2 + \nu\|\hat{\phi}^{(j)}\| \le \frac{1}{2n}\left\|\mathbf{W}_j\right\|_2^2 + \nu\|\phi^{(j)}\|_1,$$

and further it implies that

$$\begin{aligned}\frac{1}{2n}\left\|\mathbf{X}\left(\phi^{(j)} - \hat{\phi}^{(j)}\right)\right\|_2^2 + \nu\|\hat{\phi}^{(j)}\|_1 &\le \frac{1}{n}\mathbf{W}_j^T\mathbf{X}\left(\hat{\phi}^{(j)} - \phi^{(j)}\right) + \nu\|\phi^{(j)}\|_1 \\ &\le \frac{1}{n}\max_k|\mathbf{W}_j^T\mathbf{X}_k|\left(\|\hat{\phi}^{(j)}\|_1 + \|\phi^{(j)}\|_1\right) + \nu\|\phi^{(j)}\|_1.\end{aligned}$$

By Lemma 48, for some constant $c_2 > 0$ and $K_5 > 0$, we have

$$\mathrm{P}\left(\frac{1}{n}\max_{j\in\hat{\mathcal{I}},1\le k\le p}\left|\mathbf{W}_j^T\mathbf{X}_k\right| \ge \nu\right) \le K_5|\hat{\mathcal{I}}|p\exp(-c_2n^{\frac{2}{3}}\nu^2) = K_5\exp\left\{\log|\hat{\mathcal{I}}| + \log p - c_2n^{\frac{2}{3}}\nu^2\right\}.$$

By taking ν as in Corollary 24, we have

$$\mathrm{P}\left(\frac{1}{n}\max_{j\in\hat{I},1\le k\le p}\left|\mathbf{W}_j^T\mathbf{X}_k\right|\ge\nu\right)\le K_5\exp\left(-\log p\right)=K_5p^{-1}.$$

And furthermore, we have for each j,

$$\frac{1}{2n}\left\|\mathbf{X}\left(\phi^{(j)}-\hat{\phi}^{(j)}\right)\right\|_2^2\le 2\nu\|\phi^{(j)}\|_1,$$

which implies that

$$\frac{1}{n}\left\|\hat{\mathbf{W}}_{\hat{I}}-\mathbf{W}_{\hat{I}}\right\|_F^2=\sum_{j\in\hat{I}}\frac{1}{2n}\left\|\mathbf{X}\left(\phi^{(j)}-\hat{\phi}^{(j)}\right)\right\|_2^2\le 2\nu\sum_{j\in\hat{I}}\|\phi^{(j)}\|_1.$$

Finally, we use the similar proof in Section C.3 to deal with the empirical processes $n^{-1}\max_j|\boldsymbol{\varepsilon}^T\mathbf{X}_j|$ and $n^{-1}\max_{j\in\hat{I}}|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j|$. In particular, by the duality of the lasso problem (4.5),

$$\left\|\hat{\mathbf{W}}_j\right\|_2^2=\left\|\mathbf{Z}_j-\mathbf{X}\hat{\phi}^{(j)}\right\|_2^2=\min_{\mu}\left\{\left\|\mathbf{Z}_j-\mu\right\|_2^2\quad\text{s.t.}\quad\frac{1}{n}\left\|\mathbf{X}^T\mu\right\|_\infty\le\lambda\right\}\le\left\|\mathbf{Z}_j\right\|_2^2.$$

And a Hoeffding-type inequality implies that

$$\mathrm{P}\left(\frac{1}{n}\left|\boldsymbol{\varepsilon}^T\hat{\mathbf{W}}_j\right|>\lambda\middle|\hat{\mathbf{W}}_j\right)\le\exp\left(1-\frac{c_4\lambda^2n^2}{\sigma^2\|\hat{\mathbf{W}}_j\|_2^2}\right)\le\exp\left(1-\frac{c_4\lambda^2n^2}{\sigma^2\|\mathbf{Z}_j\|_2^2}\right).$$

By considering the event $\mathcal{E}_j=\left[\|\hat{\mathbf{W}}_j\|_2^2\le\|\mathbf{Z}_j\|_2^2\le n\{x+\mathrm{E}(Z_j^2)\}\right]$ for any $x>0$, the proof goes through.

C.6 Proof of Theorem 25

We start from the basic inequality that

$$\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\check{\theta}\|_2^2+\lambda\|\check{\theta}\|_1\le\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\theta^*\|_2^2+\lambda\|\theta^*\|_1,$$

which implies that

$$\frac{1}{2n}\|\mathbf{X}\check{\theta} - \mathbf{X}\theta^*\|_2^2 + \lambda\|\check{\theta}\|_1 \leq \frac{1}{n}\left(\check{\theta} - \theta^*\right)^T \mathbf{X}^T\left(\mathbf{W}\gamma^* + \varepsilon\right) + \lambda\|\theta^*\|_1.$$

The "empirical process" part can be bounded by

$$\frac{1}{n}\left|\left(\check{\theta} - \theta^*\right)^T \mathbf{X}^T\left(\mathbf{W}\gamma^* + \varepsilon\right)\right| \leq \frac{1}{n}\max_{1\leq j\leq p}\left|\mathbf{X}_j^T\left(\mathbf{W}\gamma^* + \varepsilon\right)\right| \|\check{\theta} - \theta\|_1.$$

Denote the event

$$\mathcal{T} = \left\{\frac{1}{n}\max_{1\leq j\leq p}\left|\mathbf{X}_j^T\left(\mathbf{W}\gamma^* + \varepsilon\right)\right| \leq \lambda_0 \text{ for some } \lambda_0 > 0\right\}.$$

Then on $\mathcal{T}$, for any $\lambda \geq \lambda_0$,

$$\frac{1}{2n}\|\mathbf{X}\check{\theta} - \mathbf{X}\theta^*\|_2^2 + \lambda\|\check{\theta}\|_1 \leq \lambda\|\check{\theta} - \theta^*\|_1 + \lambda\|\theta^*\|_1,$$

which further implies the slow rate bound in prediction error, i.e., $\frac{1}{2n}\|\mathbf{X}\check{\theta} - \mathbf{X}\theta^*\|_2^2 \leq 2\lambda\|\theta^*\|_1$. We conclude the proof with the following Lemma, which characterizes the scale of λ_0 and the probability that $\mathcal{T}$ holds:

Lemma 50. *Under assumption **A1**, for any $t^2 > 0$, take*

$$\lambda_0 = K_1\sigma\sqrt{\frac{t^2 + \log p}{n}} + K_2\|\gamma^*\|_1\sqrt{\frac{t^2 + \log p}{n^{2/3}}} \tag{C.12}$$

for some constants $K_1, K_2 > 0$, we have

$$\mathrm{P}\left(\frac{1}{n}\max_{1\leq j\leq p}\left|\mathbf{X}_j^T\left(\mathbf{W}\gamma^* + \varepsilon\right)\right| > \lambda_0\right) \leq C_1\exp(1 - t^2) + C_3\exp(\log p - C_2 n) \tag{C.13}$$

for some constants $C_1, C_2, C_3 > 0$.

Proof. For any $1 \leq j \leq p$,

$$\frac{1}{n}\left|\mathbf{X}_j^T\left(\mathbf{W}\gamma^* + \varepsilon\right)\right| \leq \frac{1}{n}\left|\mathbf{X}_j^T\mathbf{W}\gamma^*\right| + \frac{1}{n}\left|\mathbf{X}_j^T\varepsilon\right|.$$

We start with $n^{-1}\mathbf{X}_j^T\varepsilon$.

Since X_j is sub-Gaussian, there exists a constant $K > 0$ such that $\mathrm{E}(|X_j|^2) \leq K^2$. Consider the event $\mathcal{E}_j = \{\|\mathbf{X}_j\|_2^2 \leq 2K^2 n\}$. A Bernstein-type (Vershynin 2010) inequality implies that

$$\mathrm{P}\left(\mathcal{E}_j^C\right) = \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{ij}^2 > 2K^2\right) \leq \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{ij}^2 - \mathrm{E}(X_j^2)\right| > 2K^2 - \mathrm{E}(X_j^2)\right)$$
$$\leq c_1 \exp\left\{-c_3 n\left(2K^2 - \mathrm{E}(Z_j^2)\right)\right\} \leq c_1 \exp\left(-c_2 n\right),$$

for some $c_1, c_2, c_3 > 0$. Conditioning on that $\mathbf{X}_j = \tilde{\mathbf{X}}_j$, $n^{-1}\tilde{\mathbf{X}}_j^T\varepsilon$ is a sub-Gaussian random variable. The following Hoeffding-type inequality (Vershynin 2010) implies that

$$\mathrm{P}\left(\frac{1}{n}\left|\mathbf{X}_j^T\varepsilon\right| > x \middle| \mathbf{X}_j = \tilde{\mathbf{X}}_j\right) \leq \exp\left(1 - \frac{c_4 x^2 n^2}{\sigma^2\|\tilde{\mathbf{X}}_j\|_2^2}\right) \leq \exp\left(1 - \frac{c_4 n x^2}{2\sigma^2 K^2}\right),$$

for some constant $c_4 > 0$.

For any $t > 0$, take $x = (t^2 + \log p)^{1/2} n^{-1/2} 2^{1/2} c_4^{-1/2}\sigma K$, then

$$\mathrm{P}\left(\frac{1}{n}\left|\mathbf{X}_j^T\varepsilon\right| > x\right) \leq \mathrm{P}\left(\frac{1}{n}\left|\mathbf{X}_j^T\varepsilon\right| > x \middle| \mathcal{E}_j\right) + \mathrm{P}\left(\mathcal{E}^C\right) \leq \exp\left(1 - t^2 - \log p\right) + c_1\exp\left(-c_2 n\right).$$

So using a union bound, we have

$$\mathrm{P}\left(\max_{1\leq j\leq p}\frac{1}{n}\left|\mathbf{X}_j^T\varepsilon\right| > x\right) \leq p\exp\left(1 - t^2 - \log p\right) + p\exp\left(-cn\right) = \exp(1 - t^2) + c_1\exp(\log p - c_2 n).$$

Similarly,

$$\max_{1\leq j\leq p}|\mathbf{X}_j^T\mathbf{W}\gamma^*| \leq \max_{1\leq j\leq p}\max_{1\leq k\leq q}|\mathbf{X}_j^T\mathbf{W}_k|\|\gamma^*\|_1 = \max_{1\leq j\leq p}\max_{1\leq k\leq q}\sum_{i=1}^n|\mathbf{X}_{ij}\mathbf{W}_{ik}|\|\gamma^*\|_1.$$

By Lemma 48, we have

$$\mathrm{P}\left(\max_{1\leq k\leq q}\max_{1\leq j\leq p}\left(\mathbf{X}_j^T\mathbf{W}_k\right)^2 > n^2\varepsilon^2\right) \leq pqC_1\exp(-3C_2 n^{\frac{2}{3}}\varepsilon^2) \leq C_1\exp\left(3\log p - 3C_2 n^{\frac{2}{3}}\varepsilon^2\right).$$

Take $\varepsilon = \sqrt{\frac{t^2+\log p}{C_2 n^{2/3}}}$ for any t,

$$\mathrm{P}\left(\max_{1\leq k\leq p}\max_{1\leq j\leq p}\frac{1}{n}\left|\mathbf{X}_j^T\mathbf{W}_k\right| > \sqrt{\frac{t^2+\log p}{C_2 n^{2/3}}}\right) \leq C_1\exp\left(-t^2\right).$$

Finally, the result follows from a union bound. □

BIBLIOGRAPHY

Antoniadis, A. (2010), 'Comments on: ℓ_1-penalization for mixture regression models', *Test* **19**(2), 257–258.

Aragam, B. & Zhou, Q. (2015), 'Concave penalized estimation of sparse gaussian bayesian networks', *Journal of Machine Learning Research* **16**, 2273–2328.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008), 'Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data', *The Journal of Machine Learning Research* **9**, 485–516.

Barut, E., Fan, J. & Verhasselt, A. (2016), 'Conditional sure independence screening', *Journal of the American Statistical Association* **111**(515), 1266–1277.

Bayati, M., Erdogdu, M. A. & Montanari, A. (2013), Estimating lasso risk and noise level, *in* 'Advances in Neural Information Processing Systems', pp. 944–952.

Belloni, A., Chernozhukov, V. & Wang, L. (2011), 'Square-root lasso: pivotal recovery of sparse signals via conic programming', *Biometrika* **98**(4), 791–806.

Bickel, P. J. & Levina, E. (2008), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* **36**(1), 199–227.

Bien, J., Bunea, F. & Xiao, L. (2016), 'Convex banding of the covariance matrix', *Journal of the American Statistical Association* **111**(514), 834–845.

Bien, J., Taylor, J. & Tibshirani, R. (2013), 'A lasso for hierarchical interactions', *Annals of statistics* **41**(3), 1111.

Bien, J. & Tibshirani, R. J. (2011), 'Sparse estimation of a covariance matrix', *Biometrika* **98**(4), 807–820.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: A nonasymptotic theory of independence*, OUP Oxford.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122.

Boyd, S. & Vandenberghe, L. (2004), *Convex optimization*, Cambridge university press.

Bühlmann, P. (2013), 'Statistical significance in high-dimensional linear models', *Bernoulli* **19**(4), 1212–1242.

Bühlmann, P., Kalisch, M. & Meier, L. (2014), 'High-dimensional statistics with a view toward applications in biology', *Annual Review of Statistics and Its Application* **1**, 255–278.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cai, T., Liu, W. & Luo, X. (2011), 'A constrained ℓ_1 minimization approach to sparse precision matrix estimation', *Journal of the American Statistical Association* **106**(494), 594–607.

Campbell, F., Allen, G. I. et al. (2017), 'Within group variable selection through the exclusive lasso', *Electronic Journal of Statistics* **11**(2), 4220–4257.

Candes, E. & Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *The Annals of Statistics* pp. 2313–2351.

Chatterjee, S. & Jafarov, J. (2015), 'Prediction error of cross-validated lasso', *arXiv preprint arXiv:1502.06291* .

Choi, N. H., Li, W. & Zhu, J. (2010), 'Variable selection with the strong heredity constraint and its oracle property', *Journal of the American Statistical Association* **105**(489), 354–364.

Consortium, I. H. . et al. (2010), 'Integrating common and rare genetic variation in diverse human populations', *Nature* **467**(7311), 52–58.

Cordell, H. J. (2009), 'Detecting gene–gene interactions that underlie human diseases', *Nature Reviews Genetics* **10**(6), 392–404.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009), *Introduction to algorithms*, MIT press.

Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. (2002), 'A perspective on epistasis: limits of models displaying no main effect', *The American Journal of Human Genetics* **70**(2), 461–471.

Dalalyan, A. & Chen, Y. (2012), Fused sparsity and robust estimation for linear models with unknown variance, *in* 'Advances in Neural Information Processing Systems', pp. 1259–1267.

Dalalyan, A. S., Hebiri, M. & Lederer, J. (2017), 'On the prediction performance of the lasso', *Bernoulli* **23**(1), 552–581.

Dicker, L. H. (2014), 'Variance estimation in high-dimensional linear models', *Biometrika* **101**(2), 269.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Fan, J., Guo, S. & Hao, N. (2012), 'Variance estimation using refitted cross-

validation in ultrahigh dimensional regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(1), 37–65.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Fan, Y., Kong, Y., Li, D. & Lv, J. (2016), 'Interaction Pursuit with Feature Screening and Selection', *ArXiv e-prints* .

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. et al. (2007), 'Pathwise coordinate optimization', *The Annals of Applied Statistics* **1**(2), 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**(3), 432–441.

Gaynanova, I. (n.d.), Personal communication.

Giraud, C. (2014), *Introduction to high-dimensional statistics*, Vol. 138, CRC Press.

Hamada, M. & Wu, C. J. (1992), 'Analysis of designed experiments with complex aliasing', *Journal of quality technology* **24**(3), 130–137.

Hao, N., Feng, Y. & Zhang, H. H. (2018), 'Model selection for high-dimensional quadratic regression via regularization', *Journal of the American Statistical Association* **113**(522), 615–625.

Hao, N. & Zhang, H. H. (2014), 'Interaction screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **109**(507), 1285–1301.

Haris, A., Witten, D. & Simon, N. (2016), 'Convex modeling of interactions with strong heredity', *Journal of Computational and Graphical Statistics* **25**(4), 981–1004.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2011), *The elements of statistical learning: data mining, inference, and prediction*, Springer.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction, Second Edition*, Springer Verlag, New York.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical learning with sparsity: the lasso and generalizations*, CRC press.

Huang, J. Z., Liu, N., Pourahmadi, M. & Liu, L. (2006), 'Covariance matrix selection and estimation via penalised normal likelihood', *Biometrika* **93**(1), 85–98.

Javanmard, A. & Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression.', *Journal of Machine Learning Research* **15**(1), 2869–2909.

Jenatton, R., Audibert, J.-Y. & Bach, F. (2011), 'Structured variable selection with sparsity-inducing norms', *The Journal of Machine Learning Research* **12**, 2777–2824.

Khare, K., Oh, S., Rahman, S. & Rajaratnam, B. (2016), 'A convex framework for high-dimensional sparse Cholesky based covariance estimation', *ArXiv e-prints* .

Khare, K., Oh, S.-Y. & Rajaratnam, B. (2014), 'A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence

guarantees', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4), 803–825.

Lederer, J., Yu, L. & Gaynanova, I. (2016), 'Oracle Inequalities for High-dimensional Prediction', *ArXiv e-prints* .

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. et al. (2016), 'Exact post-selection inference, with application to the lasso', *The Annals of Statistics* **44**(3), 907–927.

Levina, E., Rothman, A. & Zhu, J. (2008), 'Sparse estimation of large covariance matrices via a nested lasso penalty', *The Annals of Applied Statistics* **2**(1), 245–263.

Lim, M. & Hastie, T. (2015), 'Learning interactions via hierarchical group-lasso regularization', *Journal of Computational and Graphical Statistics* **24**(3), 627–654.

Liu, H., Wang, L. et al. (2017), 'Tiger: a tuning-insensitive approach for optimally estimating gaussian graphical models', *Electronic Journal of Statistics* **11**(1), 241–294.

Liu, W. & Luo, X. (2012), 'High-dimensional sparse precision matrix estimation via sparse column inverse operator', *arXiv preprint arXiv:1203.3896* .

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014), 'A significance test for the lasso', *Annals of statistics* **42**(2), 413.

Lorbert, A. (2012), Alignment and supervised learning with functional neuroimaging data, PhD thesis.
URL: *http://arks.princeton.edu/ark:/88435/dsp01707957683*

Lorbert, A., Eis, D. J., Kostina, V., Blei, D. M. & Ramadge, P. J. (2010), Exploit-

ing covariate similarity in sparse regression via the pairwise elastic net., *in* 'AISTATS', Vol. 9, pp. 477–484.

Meinshausen, N. & Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* pp. 1436–1462.

Nelder, J. (1977), 'A reformulation of linear models', *Journal of the Royal Statistical Society. Series A (General)* pp. 48–77.

Ning, Y. & Liu, H. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', *Ann. Statist.* **45**(1), 158–195.

Niu, Y. S., Hao, N. & Zhang, H. H. (2018), 'Interaction screening by partial correlation', *Statistics and Its Interface* **11**(2), 317–325.

Peixoto, J. L. (1987), 'Hierarchical variable selection in polynomial regression models', *The American Statistician* **41**(4), 311–313.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009), 'Partial correlation estimation by joint sparse regression models', *Journal of the American Statistical Association* **104**(486), 735–746.

Pourahmadi, M. (1999), 'Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation', *Biometrika* **86**(3), 677–690.

Pourahmadi, M. (2013), *High-dimensional covariance estimation: with high-dimensional data*, John Wiley & Sons.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Radchenko, P. & James, G. M. (2010), 'Variable selection using adaptive nonlinear interaction structures in high dimensions', *Journal of the American Statistical Association* **105**(492), 1541–1553.

Raskutti, G., Wainwright, M. J. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over-balls', *Information Theory, IEEE Transactions on* **57**(10), 6976–6994.

Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011), 'High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence', *Electronic Journal of Statistics* **5**, 935–980.

Reid, S., Tibshirani, R. & Friedman, J. (2016), 'A study of error variance estimation in lasso regression', *Statistica Sinica* pp. 35–67.

Rigollet, P. & Tsybakov, A. (2011), 'Exponential screening and optimal rates of sparse estimation', *The Annals of Statistics* pp. 731–771.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008), 'Sparse permutation invariant covariance estimation', *Electronic Journal of Statistics* **2**, 494–515.

Schmidt, M. & Murphy, K. (2010), Convex structure learning in log-linear models: Beyond pairwise potentials, *in* 'Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics', pp. 709–716.

Shah, R. D. (2016), 'Modelling interactions in high-dimensional data with backtracking', *Journal of Machine Learning Research* **17**(207), 1–31.

She, Y., Wang, Z. & Jiang, H. (2018), 'Group regularized estimation under structural hierarchy', *Journal of the American Statistical Association* **113**(521), 445–454.

Shojaie, A. & Michailidis, G. (2010), 'Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs', *Biometrika* **97**(3), 519–538.

Städler, N., Bühlmann, P. & van de Geer, S. (2010), 'ℓ_1-penalization for mixture regression models (with discussion)', *Test* **19**, 209–285.

Sun, T. & Zhang, C.-H. (2010), 'Comments on: ℓ_1-penalization for mixture regression models', *Test* **19**(2), 270–275.

Sun, T. & Zhang, C.-H. (2012), 'Scaled sparse linear regression', *Biometrika* **99**(4), 879–898.

Sun, T. & Zhang, C.-H. (2013), 'Sparse matrix inversion with scaled lasso', *The Journal of Machine Learning Research* **14**(1), 3385–3418.

Taylor, J. & Tibshirani, R. (2017), 'Post-selection inference for ℓ_1-penalized likelihood models', *Canadian Journal of Statistics* .

Thanei, G.-A., Meinshausen, N. & Shah, R. D. (2016), 'The xyz algorithm for fast interaction search in high-dimensional data', *ArXiv e-prints* .

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., Wasserman, L. et al. (2018), 'Uniform asymptotic inference and the bootstrap after model selection', *The Annals of Statistics* **46**(3), 1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* **111**(514), 600–620.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for non-differentiable minimization', *Journal of Optimization Theory and Applications* **109**(3), 475–494.

Turlach, B. A. (2004), 'Discussion of "least angle regression" by efron et al', *arXiv preprint math/0406472* .

Van de Geer, S. A. & Bühlmann, P. (2009), 'On the conditions used to prove oracle results for the lasso', *Electronic Journal of Statistics* **3**, 1360–1392.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. et al. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**(3), 1166–1202.

Vershynin, R. (2010), 'Introduction to the non-asymptotic analysis of random matrices', *arXiv preprint arXiv:1011.3027* .

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)', *Information Theory, IEEE Transactions on* **55**(5), 2183–2202.

Wu, J., Devlin, B., Ringquist, S., Trucco, M. & Roeder, K. (2010), 'Screen and clean: a tool for identifying interactions in genome-wide association studies', *Genetic epidemiology* **34**(3), 275–285.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. (2009), 'Genome-wide association analysis by lasso penalized logistic regression', *Bioinformatics* **25**(6), 714–721.

Wu, W. B. & Pourahmadi, M. (2003), 'Nonparametric estimation of large covariance matrices of longitudinal data', *Biometrika* **90**(4), 831–844.

Yan, X. & Bien, J. (2015), 'Hierarchical sparse modeling: A choice of two regularizers', *arXiv preprint arXiv:1512.01631* .

Yu, G. (2017), *natural: Estimating the Error Variance in a High-Dimensional Linear Model*. R package version 0.9.0.
URL: *https://CRAN.R-project.org/package=natural*

Yu, G. & Bien, J. (2017*a*), 'Estimating the error variance in a high-dimensional linear model', *ArXiv e-prints* .

Yu, G. & Bien, J. (2017*b*), 'Learning local dependence in ordered data', *Journal of Machine Learning Research* **18**(42), 1–60.

Yuan, M. (2010), 'High dimensional inverse covariance matrix estimation via linear programming', *The Journal of Machine Learning Research* **11**, 2261–2286.

Yuan, M., Joseph, V. R. & Zou, H. (2009), 'Structured variable selection and estimation', *The Annals of Applied Statistics* pp. 1738–1757.

Yuan, M. & Lin, Y. (2007), 'Model selection and estimation in the gaussian graphical model', *Biometrika* **94**(1), 19–35.

Zhang, C.-H. & Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

Zhang, T. & Zou, H. (2014), 'Sparse precision matrix estimation via lasso penalized d-trace loss', *Biometrika* **101**(1), 103–120.

Zhao, P., Rocha, G. & Yu, B. (2009), 'The composite absolute penalties family for grouped and hierarchical variable selection', *The Annals of Statistics* **37**(6A), 3468–3497.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *The Journal of Machine Learning Research* **7**, 2541–2563.

Zhou, Y., Jin, R. & Hoi, S. (2010), Exclusive lasso for multi-task feature selection, *in* 'International conference on artificial intelligence and statistics', pp. 988–995.

Zimmerman, D. L. & Nunez-Anton, V. A. (2009), *Antedependence models for longitudinal data*, CRC Press.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.